A CORRELATED RANDOM EFFECTS HURDLE MODEL FOR EXCESS ZEROS WITH CLUSTERED DATA BASED ON BLUP (REMQL) ESTIMATION

by

Sung Hee Kim

M.S., Rutgers, The State University of New Jersey, 2005

B.S., Dongguk University, South Korea, 2002

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Sung Hee Kim

It was defended on

April 01, 2011

and approved by

Roslyn A. Stone, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Chung-Chou H. Chang, PhD, Associate Professor, Departments of Medicine and Biostatistics, School of Medicine and Graduate School of Public Health, University of Pittsburgh

Kevin H. Kim, PhD, Associate Professor, Department of Psychology, School of Education, University of Pittsburgh

Michael J. Fine, MD, Professor, Department of Medicine, School of Medicine, University of Pittsburgh

Sati Mazumdar, PhD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Roslyn A. Stone, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

A CORRELATED RANDOM EFFECTS HURDLE MODEL FOR EXCESS ZEROS WITH CLUSTERED DATA BASED ON BLUP (REMQL) ESTIMATION

Sung Hee Kim, PhD

University of Pittsburgh, 2011

Community-acquired pneumonia (CAP) is a common, costly, and fatal illness; more than four million episodes occur in the United States each year. Providing quality and cost-effective care for CAP has an important implication in public health. Since inpatient treatment costs 20 times as much as outpatient treatment, and the costs of hospitalization drive inpatient costs, reducing length of stay (LOS) for patients with CAP may substantially reduce medical care costs and improve the effectiveness of health utilization. A potentially useful metric of efficiency is bed days, defined as zero for outpatients and LOS for inpatients, where LOS is the difference between discharge and admission dates. A surrogate for hospitalization costs, bed days, has problematic statistical properties (i.e., excess zeros and possible overdispersion). In multi-site studies, we also need to account for possible clustering by site.

Researchers used finite mixture (FM) models or zero-inflated (ZI) models for bed days, assuming that valid zeros occur in both component distributions. However, the hurdle (H) model presumes all zero bed days are from outpatients. The H model has been extended to include correlated random effects in a generalized linear mixed model (GLMM) framework previously. Maximum Likelihood (ML) estimation is one of the most common approaches for estimating GLMMs. To avoid the intensive computing, convergence problems, and biased estimates of variance components associated with ML, we implemented best linear unbiased prediction (BLUP)-type estimation with restricted maximum quasi-likelihood (REMQL) of variance components in the correlated random effects H model. Several simu-

lation studies validate this approach. We also applied the proposed random effects H model to the Emergency Department Community Acquired Pneumonia (EDCAP) study, a 32-site cluster-randomized trial to assess the effect of implementing medical practice guidelines on two aspects of care, e.g., admission and LOS. This allowed us to investigate whether the distribution of bed days varies by intervention arm (site-level) and the risk status (patient-level) among low risk patients with pneumonia. This appropriate modeling of bed days may facilitate identification of predictors of costly hospitalizations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

First, I would like to express my sincere appreciation to my dissertation advisor, Dr. Roslyn A. Stone, for her insight, direction, constant encouragement, and support. I also would like to thank to my committee members: Dr. Chung-Chou H. Chang, for her valuable discussions on estimation; Dr. Kevin H. Kim, for his input and help in simulation studies; Dr. Michael J. Fine, for graciously allowing me to access the EDCAP data, which motivated the main idea of my dissertation from the Center for Health Equity Research and Promotion (CHERP); and Dr. Sati Mazumdar, for her indispensable lecture on mixed models. I am grateful to all these individuals who encouraged me and enlightened me during the process of my dissertation.

I want to thank the Biostatistics Department and CHERP, which provided financial support through my Ph.D study. Finally, I dedicate this dissertation to my family and friends, whose love, confidence sustained me in my doctoral journey.

1.0 INTRODUCTION

Community-acquired pneumonia (CAP) is a common, costly, and often fatal illness with more than four million episodes in the United States each year (Hsu *et al.*, 2010[18]). CAP, also a precreation of severe sepsis, leads to multi-organ system failure, particularly respiratory distress and shock (Renaud *et al.*, 2009[38]). Since CAP is associated with a high rate of morbidity, mortality, and cost of care (Yealy *et al.*, 2004[50]), we need to pay attention to how we can provide qualitative and cost-effective care for CAP. The direct medical care costs of treating patients with pneumonia are almost $10 billion per year, with the cost of inpatient treatment being 20 times as much as outpatient treatment (Niederman *et al.*, 1998[31]; Fine *et al.*, 2000[10]). Because inpatient cost consists primarily of the cost of hospitalizations, appropriately reducing the admission rate of low risk patients and the length of stay (LOS) for inpatients with CAP may contribute substantially to medical care cost savings and efficient health care utilization (Fine *et al.*, 2000[10]).

The Emergency Department Community Acquired Pneumonia (EDCAP) study is a 32-site study to assess the effectiveness and safety of three guideline implementation strategies of increasing intensity (low intensity, moderate intensity, and high intensity) to increase the proportion of low risk patients who are treated as outpatients. Low risk patients were identified using a validated measure of pneumonia severity, the Pneumonia Severity Index (PSI) (Yealy *et al.*, 2005[51]). The low intensity intervention occurred at eight sites, while the moderate and high intensity interventions each occurred by 12 sites. The sample study consisted of 3,219 patients from Connecticut and Pennsylvania with clinical and radiographic evidence of pneumonia. Fifty-nine percent (N=1,901) were low risk patients (i.e., PSI$\leq$3 without hypoxemia), and 41% (N=1,318) were high risk patients (i.e., PSI>3 or with hypoxemia). Using a 3-level logistic regression with the levels defined by patients, medical

providers, and sites, more low risk patients were treated as outpatients in the moderate intensity and high intensity interventions than in the low intensity intervention (low intensity intervention, 37.5%; moderate intensity intervention, 61.0%; high intensity intervention, 61.9%).

Figure 1 shows the empirical distribution of bed days for each PSI risk class by intervention arm among low risk patients. The spikes at zero bed days represent outpatients; the prevalence of outpatient care decreases with increasing risk class within each intervention arm. The distributions for patients at the moderate and high intensity intervention sites are similar, and both are right skewed. Both the moderate and high intensity intervention sites had more outpatients and fewer inpatient bed days than the low intensity intervention sites.

For measures of efficiency of care, we can look at the probability of outpatient treatment as the initial EDCAP study did, or we can investigate inpatient LOS (Yau *et al.*, 2003[49]). One alternative measure of efficiency that includes both components is "bed days", defined as zero for outpatients and LOS for inpatients, where LOS is the difference between discharge and admission dates (Wang *et al.*, 2002[44]). We also treat LOS<24 hours as one bed day for inpatients, because the cost of care for inpatients is much more expensive than that for outpatients, even though the patient receives less than one day of hospital care. Among 3,219 in the EDCAP cohort, 24 patients have missing information on bed days, 1,302 patients are high risk, and 16 died within 30 days. Our analyses included 1,877 low risk patients (1,061 outpatients and 816 inpatients) who had enrolled in the EDCAP trial and remained alive at 30 days (Figure 2).

In similar studies, researchers have used Poisson or zero-truncated Poisson regression and negative binomial or zero-truncated negative binomial regression for overdispersed cases to model for inpatient LOS or bed days (Page *et al.*, 2002[32]; Xie and Aickin, 1997[48]; Brown *et al.*, 2003[5]; Lee *et al.*, 2003[22]). However, these methods do not account for excess zeros from outpatients.

Finite mixture (FM) models, zero-inflated (ZI) models, and hurdle (H) models allow for excess zeros in the data. In a g-component Poisson mixture model, the number of components must be estimated (Schlattmann *et al.*, 1996[41];Wang *et al.*, 1996[46]). For bed days, the number of components is known to be two, i.e. inpatients and outpatients. Although the

ZI and H models accommodate counts with excess zeros (Cunningham and Lindenmayer, 2005[6]; Min and Agresti, 2002[28]; Ridout *et al.*, 1998[39]; Welsh *et al.*, 1996[47]), we will not consider ZI models here since the ZI model allows zeros to occur in both component distributions. H models have been used in economic applications and health care services (Arulampalam and Booth, 1997[2]; Gurmu, 1998[14]; Pohlmeier and Ulrich, 1995[35]). An H model is a two-component mixture with a binomial part (probability of passing the "hurdle") and a Poisson or negative binomial part. An H model is appropriate for the outcome of bed days because all patients are at risk for hospitalization when they present to a site (hospital), but zero bed days occur only among outpatients.

In multi-site studies, the H model has been extended to account for clustering by site because patients collected from the same site are correlated (Lee *et al.*, 2007[21]). Ignoring the dependency within sites may result in underestimated variance estimates of the coefficients (Song, 2005[42]). Generalized linear mixed models (GLMMs) frequently have been used to accommodate such the clustering; their flexibility allows for a discrete and non-normally distributed outcome and the incorporation of random effects. Maximum Likelihood (ML) commonly is used to estimate parameters from the marginal likelihood, with numerical approximations to integrate out the random effects. Min and Agresti (2005)[29] presented a correlated random effects H model using the ML approach. Although the ML approach is efficient, it involves intensive computing and may not converge. In addition, the ML approach can give biased estimates of variance components for random effects. An alternative method to estimate variance components in the GLMM setting, a best linear unbiased prediction (BLUP)-type estimation with an approximate restricted maximum quasi-likelihood (REMQL), requires less integration and produces less biased estimates of variance components relative to ML (McGilchrist, 1994[25]; McGilchrist and Yau, 1995[26]). Although the BLUP (REMQL) approach has been used to estimate random effects in FM and ZI models, it has not been implemented for the random effects H model. In this dissertation, we will address several statistical issues based on simulation studies:

1. In the correlated random effects H model, does the proposed BLUP (REMQL) estimation give unbiased and consistent estimators of regression coefficients similar to the ML estimation?

2. Does the correlated random effects H model with the proposed BLUP (REMQL) estimation produce less biased estimates of the variance components relative to ML estimation?
3. In terms of the computational effort, is the proposed BLUP (REMQL) more efficient than the ML?

Our proposed BLUP (REMQL) approach will be applied to the EDCAP data to illustrate the method. In addition, we can identify unusual sites using the random effects prediction from the proposed BLUP (REMQL) estimation procedure.

The organization of this dissertation is as follows. In chapter 2, we describe statistical models with key properties in the analysis of bed days and introduce the BLUP (REMQL) estimation in the GLMM setting, which is implemented in the H model. Chapter 3 presents the proposed BLUP (REMQL) estimation in the Poisson H model with correlated random effects. In this chapter, we also analyze the EDCAP data and show simulation studies. The proposed BLUP (REMQL) estimation approach in the negative binomial H model with correlated random effects, as well as the EDCAP application and simulation studies, are presented in Chapter 4. Chapter 5 concludes with a discussion.

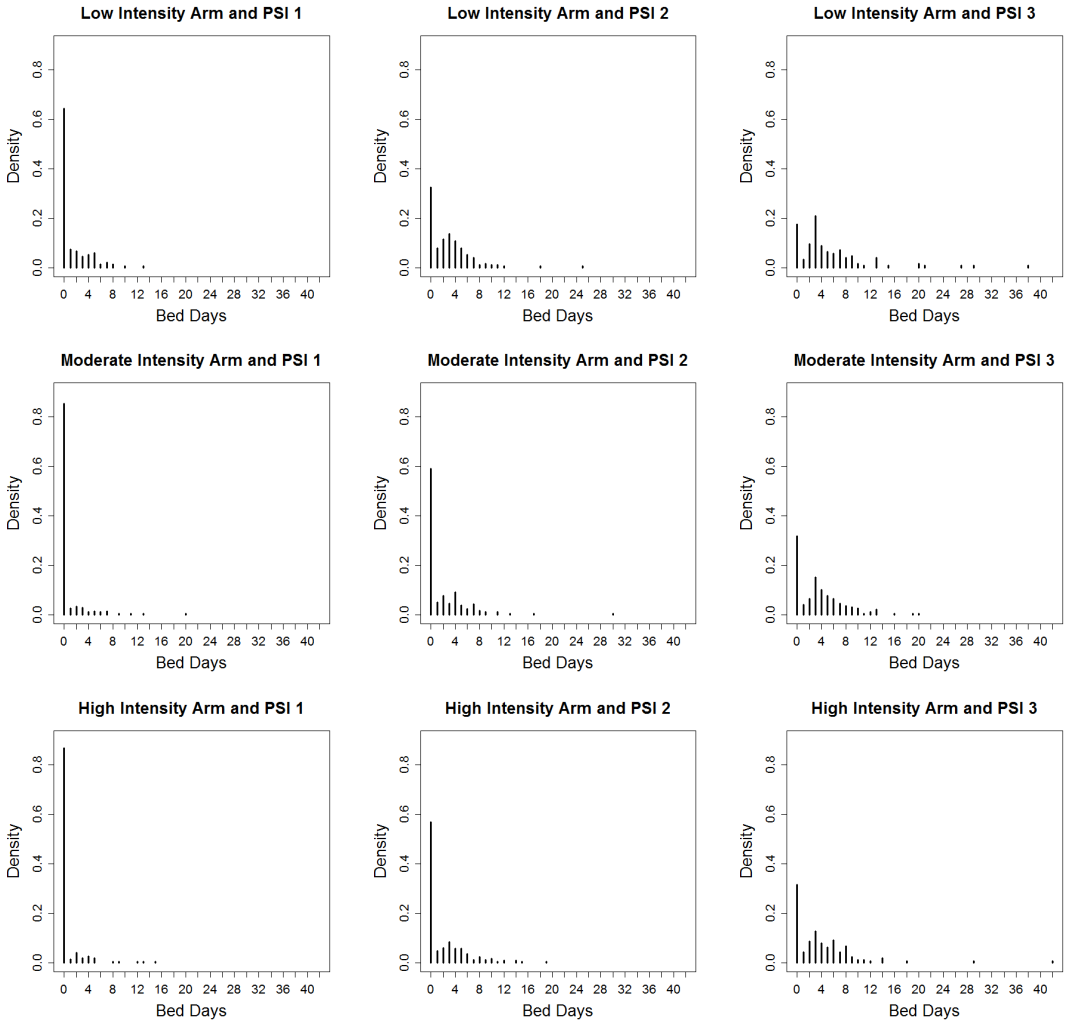

Figure 1: Bed days by intervention arm and PSI risk class in the EDCAP study

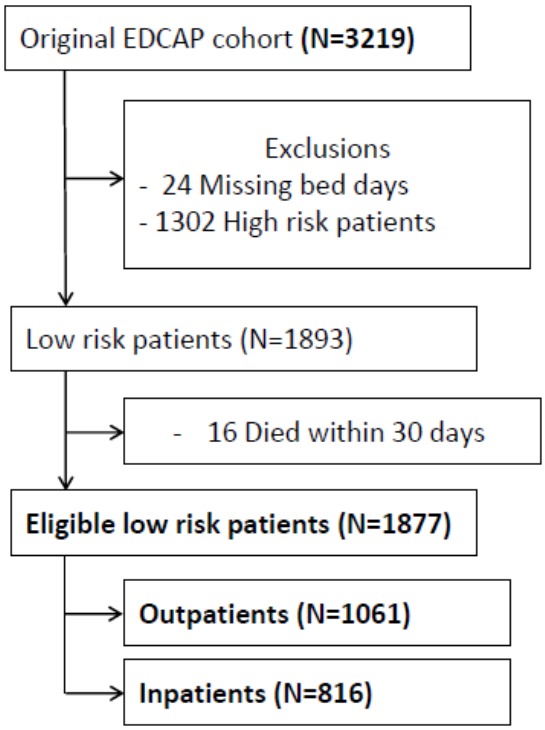

Figure 2: Original EDCAP cohort and eligible low risk patients

2.0 REVIEW OF LITERATURE

In the analysis of bed days, finite mixture (FM) models, zero-inflated (ZI) models, and hurdle (H) models have been used to account for excess zeros and often skewed distributions. In this chapter, we will first describe FM models, ZI models, and H models in detail. Then, we will summarize the properties of each model and explain advantages of the H model for modeling bed days. In addition, we will discuss the extension of the H model to account for clustering by site. Finally, in Chapters 3 and 4, we will introduce the BLUP (REMQL) estimation in the GLMM setting as an alternative to the ML in the correlated random effects H model setting.

2.1 FINITE MIXTURE MODELS

The FM model was defined and summarized well in the textbook *Finite Mixture Models* (McLachlan and Peel, 2001[27]). The FM model consists of a mixing proportion (or a membership probability) and a finite number of distributions. For the simple approach, a Poisson FM model with only a covariate adjusted component distribution has been used for a disease map construction by Schlattmann (1996)[41] and for a daily seizure count and for a number of relevant colonies of salmonella by Wang *et al.* (1996)[46]. However, in many application areas, the FM model with the covariate adjusted mixing probability is often used; for example, Dalrymple *et al.* (2003)[7] used the Poisson FM model withn analyzing sudden infant death syndrome (SIDS) data. Hence, we will give a general model, which allows for the covariate adjusted mixing probability and the covariate adjusted component distribution. Because the Poisson FM model, with its flexibility of covariate adjustment,

can be simply generalized into the FM model with the other distribution, we will present the Poisson FM model here. Let Y_i $(i = 1, 2, ..., n)$ represent the bed days of the patient i when n is the total number of patients. Suppose Y_i comes from a mixture of gth component Poisson distribution, f_k with mean $\mu_{k,i}$, where k indicates the kth subgroup . Then, the Poisson FM model is

$$P(Y_i = y_i) = \sum_{k=1}^{g} p_k f_k(y_i) = \sum_{k=1}^{g} p_k \frac{\mu_{k,i}^{y_i}}{y_i} exp(-\mu_{k,i}) \tag{2.1}$$

where p_k indicates the mixing probability, which denotes the probability of the patient belonging to the kth subgroup, that sum to one; that is,

$$0 < p_k < 1 \qquad \text{and} \qquad \sum_{k=1}^{g} p_k = 1.$$

Using logit and log-linear links to model p_k and $\mu_{k,i}$, respectively, gives:

$$\text{logit}(p_k) = \log(\frac{p_k}{1 - p_k}) = \boldsymbol{w}_{k,i}^T \boldsymbol{\alpha}_k \qquad \text{and} \qquad \log(\mu_{k,i}) = \boldsymbol{x}_{k,i}^T \boldsymbol{\beta}_k \tag{2.2}$$

where $\boldsymbol{w}_{k,i}$ and $\boldsymbol{x}_{k,i}$ indicate vectors of covariates for ith patient corresponding to regression coefficients $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ in kth subgroup, respectively. Note that g homogeneous subgroups can be affected by different covariates. In practice, the number of components g in the FM model is not observed and needs to be estimated from the data.

2.2 ZERO-INFLATED MODELS

The ZI model, proposed by Lambert (1992)[20] for count data with excess zeros, can be considered a mixture of a Bernoulli distribution at zero and a count distribution. Lambert's Poisson ZI model was used to model the data with excess zeros in many applications (Ridout *et al.*, 1998[39]; Böhning *et al.*, 1999[3]; Quintero, 2007[36]). The ZI model is one of several choices for modeling ZI count data (Welsh *et al.*, 1996[47]; Min and Agresti, 2002[28]; Cunningham and Lindenmayer, 2005[6]). Lee *et al.* (2004)[23] also used the Poisson ZI and the negative binomial ZI model when studying the sensitivity of score tests for zero-inflation with two applications (occupational injury count data and a set of pancreas disorder data). Gurmu (1996)[15] presented a semiparametric estimation approach for the ZI model to analyze recreational boating trips, and Pardoe (2003)[33] suggested a Bayesian approach in the ZI model to evaluate the impact of objective, sensory descriptors and price on the choice of wine. Based on Lambert's Poisson ZI model, the Poisson ZI model for bed days Y_i of patient i $(i = 1, 2, ..., n)$ can be written as

$$\begin{aligned} P(Y_i = 0) &= p_{ZI,i} \cdot 1 + (1 - p_{ZI,i}) \cdot f(0) = p_{ZI,i} + (1 - p_{ZI,i}) \cdot exp(-\mu_i) \\ P(Y_i = y_i) &= p_{ZI,i} \cdot 0 + (1 - p_{ZI,i}) \cdot f(y_i) = (1 - p_{ZI,i}) \cdot \frac{\mu_i^{y_i}}{y_i} exp(-\mu_i) \quad y_i = 1, 2, \dots \end{aligned} \tag{2.3}$$

where $p_{ZI,i}$ is the conditional probability of having zero bed days under the assumption that patient i is not at risk for hospitalization, and f indicates Poisson distribution with mean μ_i. Then, we can model parameters $(p_{ZI,i},\ \mu_i)$ using logit and log-linear links by

$$\text{logit}(p_{ZI,i}) = \boldsymbol{w}_i^T \boldsymbol{\alpha} \qquad \text{and} \qquad \log(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} \tag{2.4}$$

where $\boldsymbol{w}_i$ and $\boldsymbol{x}_i$ indicate vectors of covariates with respect to regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. In the Poisson ZI model, we assume that zero bed days are from both patients not at risk (structural zeros) and patients at risk (sampling zeros) for hospitalization (Rose *et al.*, 2006[40]). Contrary to the FM model, the number of components in the Poisson ZI model is known as two. The Poisson ZI model can also be regarded as a special case of the two-component FM model (Dalrymple, 2003[7]). For example, one component is taken mass 1, and the other component is taken Poisson distribution at $y_i = 0$. To account for heterogeneity

in addition to excess zeros, replacing f with a negative binomial distribution instead of a Poisson distribution produces the negative binomial ZI model (Greene, 1994[12]).

2.3 HURDLE MODELS

Another statistical model for bed days is the Poisson H model developed by Mullahy (1986)[30]. The Poisson H model can be regarded as a two-part model (Heilbron, 1994[17]). The first part is the binary response model that measures whether the response falls at or above zero, and the second part is the zero-truncated Poisson model that explains the responses above zero. Mullahy's Poisson H model was used in economic applications (Pohlmeier and Ulrich, 1995[35]) and Medicaid utilization (Gurmu, 1998[14]). Arulampalam and Booth (1997)[2] employed the negative binomial H model for estimating the number of work-related training events. In addition to Lambert's Poisson ZI model (Welsh *et al.*, 1996[47]; Ridout *et al.*, 1998[39]; Min and Agresti, 2002[28]; Cunningham and Lindenmayer, 2005[6]), the Poisson H or negative binomial H models have been suggested for modeling count data with excess zeros. Gurmu (1997)[13] also suggested applying the semiparametric approach for the Poisson H or negative binomial H models for Medicaid utilization, and Pardoe (2003)[33] worked on a Bayesian approach for both the ZI and H models. Mullahy's Poisson H model for bed days Y_i of patient i $(i = 1, 2, ..., n)$ can be defined as

$$\begin{aligned} P(Y_i = 0) &= p_{H,i} \cdot 1 + (1 - p_{H,i}) \cdot 0 = p_{H,i} \\ P(Y_i = y_i) &= p_{H,i} \cdot 0 + (1 - p_{H,i}) \cdot f(y_i)/(1 - f(0)) \\ &= (1 - p_{H,i}) \cdot \frac{\mu_i^{y_i}}{y_i} exp(-\mu_i)/(1 - exp(-\mu_i)), \qquad y_i = 1, 2, \ldots \end{aligned} \tag{2.5}$$

where $p_{H,i}$ is the conditional probability of passing the hurdle given that patient i is at risk for hospitalizations, and f indicates the Poisson distribution with mean μ_i. Then, we can model parameters ($p_{H,i}$, μ_i) using logit and log-linear links (2.4) like those described in the Poisson ZI model. In the Poisson H model, all patients are considered at risk for hospitalization. Hence, the Poisson H model does not have structural zeros, which means that all zero bed days are only from patients at risk for hospitalization. The Poisson H model presumes that,

similar to the ZI model, the number of components is known. Like the Poisson ZI model, the Poisson H model can be extended to the negative binomial H model by relaxing the assumption that the mean and variance are equal.

2.4 MODEL DECISION

As we reviewed several potential models (FM, ZI, H) for the analysis of bed days, we recognized the need to consider the study design and outcome properties before choosing our specific model. If we assume the patients are from several different populations, the finite mixture model is very attractive (Min and Agresti, 2002[28]). For example, the FM model, such as the three-component mixture model, would be a good choice for modeling if we can assume the patients are from three subgroups. In the FM model, overestimating the number of components can cause a lack of model fit. Unlike the FM model, the zero-inflated model is more suitable for handling zero inflation when the observed zeros are greater than would be expected in a particular distribution (Min and Agresti, 2005[29]). Moreover, the ZI model can treat both structural and sampling zeros. Fitting ZI model components simultaneously complicates estimation and interpretation (Kuhnert *et al.*, 2005[19]). By contrast, the H model can be fitted separately, which leads to computational merits. The H model assumes all patients are at risk for hospitalization. In our EDCAP study, we assumed that all patients were at risk for hospitalization (e.g., observed zeros arise only from patients at risk for hospitalization) when they visited the site (hospital) at emergency departments. Hence, we consider the H model more appropriate than the other models (FM, Z). However, this does not suggest that the H model is always better than the other models. The model has to be chosen to accommodate specific study design and outcome properties.

2.5 EXTENDED HURDLE MODEL WITH CORRELATED RANDOM EFFECTS

Patients from the same site (hospital) are likely to be correlated. Ignoring the clustering by site can produce biased inferences. To account for clustering by site, Min and Agresti (2005)[29] extended the Poisson Hurdle (H) model in (2.5) to include random effects by incorporating random effects into both the logistic and the log-linear parts of the H model (Min and Agresti, 2005[29]). This suggests that we use generalized linear mixed models (GLMM) not only to address excess zeros but also to incorporate the clustering. Let Y_{ij} $(i = 1, 2, ..., m; j = 1, 2, ..., n_i)$ be the number of bed days of patient j at site i, where m is the number of sites and n_i is the number of patients at site i; the total number of patients is $n = \sum_{i=1}^{m} n_i$. Then, the Poisson H model with random effects is:

$$\begin{aligned} P(Y_{ij} = 0) &= p_{H,ij}, \\ P(Y_{ij} = y_{ij} | y_{ij} > 0) &= (1 - p_{H,ij}) \cdot \frac{f(y_{ij})}{1 - f(0)} \\ &= (1 - p_{H,ij}) \frac{e^{-\mu_{ij}} \mu_{ij}^{y_{ij}} / y_{ij}!}{1 - e^{-\mu_{ij}}}, \end{aligned} \tag{2.6}$$

where $p_{H,ij}$ indicates the conditional probability of not passing the hurdle (i.e., not being hospitalized) given the jth patient at site i is at risk for hospitalization, and μ_{ij} is the mean of the underlying Poisson distribution. Note that the probability $(p_{H,ij})$ can be modeled by the logistic regression and $\frac{f(y_{ij})}{1-f(0)}$ can be regarded as a truncated Poisson distribution. In the regression setting, both $\text{logit}(p_{H,ij})$ and $\log(\mu_{ij})$ are assumed to depend on linear functions of the covariates. Following Wang's notation for the two-component mixture model (Wang *et al.*, 2007[45]), we can define the linear predictors ξ_{ij} and η_{ij} by

$$\begin{aligned} \text{logit}(p_{H,ij}) &= \xi_{ij} = \boldsymbol{w}_{ij}^T \boldsymbol{\alpha} + \boldsymbol{u_i} \\ \log(\mu_{ij}) &= \eta_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{v_i} \end{aligned} \tag{2.7}$$

where $\boldsymbol{w}_{ij}$ and $\boldsymbol{x}_{ij}$ are vectors of covariates for the logistic and the Poisson distribution, respectively, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the corresponding vectors of the coefficients. Let random

effect vectors $\boldsymbol{u_i}$ and $\boldsymbol{v_i}$ be such that $\boldsymbol{u} = (\boldsymbol{u_1}, ..., \boldsymbol{u_m})$ and $\boldsymbol{v} = (\boldsymbol{v_1}, ..., \boldsymbol{v_m})$, respectively. Then, $\boldsymbol{r}_i^T = (\boldsymbol{u}_i, \boldsymbol{v}_i)^T$ is assumed to be distributed as $N(\boldsymbol{0}, \boldsymbol{A}_i(\boldsymbol{\phi}))$, where $\boldsymbol{\phi} = (\sigma_u, \sigma_v, \rho)$.

The correlation between the binomial and count components is represented in the covariance matrix $\boldsymbol{A}_i(\boldsymbol{\phi})$, where

$$\boldsymbol{A}_i(\boldsymbol{\phi}) = \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}, \qquad i = 1, 2, ..., m, \tag{2.8}$$

when ρ denotes a bivariate correlation between the random effects. In the case of uncorrelated random effects, $\rho = 0$ in the covariance matrix $(\boldsymbol{A}_i(\boldsymbol{\phi}))$, and $\boldsymbol{u}$ and $\boldsymbol{v}$ are assumed to be independent and distributed as N(0, $\sigma_u^2 \mathrm{I}_m$) and N(0, $\sigma_v^2 \mathrm{I}_m$), respectively, where I_m denotes an $m \times m$ identity matrix. The logistic regression part and Poisson regression part can be estimated separately. However, the estimation has to be done jointly in the correlated random effects.

2.6 ESTIMATION IN THE CORRELATED RANDOM EFFECTS HURDLE MODEL

Min and Agresti (2005)[29] developed the H model with correlated random effects using a maximum likelihood (ML) estimation approach, which is the most common approach in the GLMM. A Gauss-Hermite (GH) quadrature (Fahrmeir and Tutz, 2001[8]; McCulloch and Searle, 2001[24]) was used to approximate the integral for random effects by a finite sum, and an approximate version of a Fisher scoring method (Green, 1984[11];Raudenbush *et al.*, 2000[37]), which is an iterative method to obtain the ML estimates when the Fisher information matrix is not a closed form, was adapted. They also suggested Aitkin's non-parametric maximum likelihood (NPML) approach to prevent the misspecification of the random effect distribution (Aitkin, 1999[1]). However, because the ML approach requires solving a complex form of integration with respect to the random effects' distribution, it incorporates several approximate methods, such as the GH quadrature, the adaptive Gauss-Hermite quadrature (AGQ), the Monte Carlo EM algorithm, the Markov chain Monte Carlo

(MCMC), and the Laplace approximations (Fahrmeir and Tutz, 2001[8]; McCulloch and Searle, 2001[24]). Hence, ML can have convergence problems. The difficulty in evaluating the marginal likelihood from the ML approach led to another approach: best linear unbiased prediction (BLUP)-type estimation with an approximate restricted maximum quasi-likelihood (REMQL) for variance components. In principle, the BLUP (REMQL) approach is very similar to the peneralized quasi-likelihood (PQL) approach (Breslow and Clayton, 1993[4]). The BLUP (REMQL) approach, which is computationally less complicated than the ML approach, has been popularly used in the FM or ZI models with random effects. The BLUP (REMQL) approach also provides predicted random effects, which may be considered as site efficiency indicators, thereby allowing the identification of unusual sites. Therefore, we implement a BLUP (REMQL) approach to fit the GLMM in the correlated random effects H model.

2.7 BLUP (REMQL) ESTIMATION APPROACH

The BLUP (REMQL) estimation in the GLMM was proposed by McGilchrist (1994)[25] and is explained in detail by McGilchrist and Yau (1995)[26]. This approach is based on a BLUP estimation for fixed effects and an approximate REMQL estimation for variance components of random effects, which are derived from linking the GLMM with a normal error model. We will summarize this approach following the work of McGilchrist. In a normal error model, a response vector $\boldsymbol{y}$ with n observations can be expressed as

$$\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{e}, \qquad \boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zr} \tag{2.9}$$

where $\boldsymbol{X}$ is an $n \times p$ design matrix for fixed component, $\boldsymbol{Z}$ is an $n \times m$ design matrix for a random component, $\boldsymbol{\beta}$ is an unknown parameter vector with dimension p for a fixed component, $\boldsymbol{r}$ is an unknown parameter vector for a random component, $\boldsymbol{e}$ is distributed as $\mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{D})$, and $\boldsymbol{D}$ is a known symmetric matrix of dimension n. The component $\boldsymbol{Zr}$ may be partitioned into

$$\boldsymbol{r}^T = [\boldsymbol{r}_1^T, \boldsymbol{r}_2^T, ..., \boldsymbol{r}_m^T], \qquad \boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, ..., \boldsymbol{Z}_m]$$

where $\boldsymbol{r}_i$ is independently distributed as $\mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{A}_i(\boldsymbol{\phi}))$, $\boldsymbol{\phi} = (\sigma_u, \sigma_v, \rho)$, and the covariance component matrix component $\boldsymbol{A}_i(\boldsymbol{\phi})$ is

$$\boldsymbol{A}_i(\boldsymbol{\phi}) = \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}, \qquad i = 1, 2, ..., m.$$

Then, the BLUP procedure to estimate $\boldsymbol{\beta}$, $\boldsymbol{r}$, σ^2, and $\boldsymbol{\phi}$ maximizes the BLUP-type loglikelihood of $\boldsymbol{y}$ and $\boldsymbol{r}$, which can be expressed as

$$\ell(\boldsymbol{y}, \boldsymbol{r}) = \ell_1(\boldsymbol{y}|\boldsymbol{r}) + \ell_2(\boldsymbol{r}) \tag{2.10}$$

where ℓ_1 is the loglikelihood for $\boldsymbol{y}$ conditional on $\boldsymbol{r}$, and ℓ_2 is the loglikelihood for the non-observable $\boldsymbol{r}$. The estimating equations for $\boldsymbol{\beta}$ and $\boldsymbol{r}$ are:

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{D}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{D}^{-1}\boldsymbol{Z} + \boldsymbol{A}^{-1} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{r}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{D}^{-1}\boldsymbol{y} \end{pmatrix} \tag{2.11}$$

where $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{r}}$ are the BLUP estimates, and $\boldsymbol{A} = [\boldsymbol{A}_1(\phi), ..., \boldsymbol{A}_m(\phi)]$ denotes a block diagonal matrix.

To reduce bias of the estimated BLUP variance components, Harville (1977)[16], Thompson (1980)[43], and Fellner (1987)[9] derived ML and REML estimators of variance components from BLUP estimators in normal error models as summarized by McGilchrist (1994)[25]. In the GLMM, the response vector $\boldsymbol{y}$ is not necessarily normally distributed, so that $\ell_1(\boldsymbol{y}|\boldsymbol{r})$ is not of normal form in general. In this case, McGilchrist (1994)[25] suggested using an approximate asymptotic normal distribution with mean $\boldsymbol{\beta}$ and $\boldsymbol{r}$ and a variance matrix given by the information matrix for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{r}}$, which is

$$\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{Z}^T \end{pmatrix} \boldsymbol{B} \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Z} \end{pmatrix},$$

where $\boldsymbol{B} = -E(\partial^2\ell_1/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^T)$ when ℓ is approximately quadratic in $\boldsymbol{\beta}$ and $\boldsymbol{r}$. Hence, the BLUP-type loglikelihood (2.10) is

$$\ell^* = \ell_1^* + \ell_2$$

where

$$\begin{aligned}
\ell_1^* &= \text{constant} - \frac{1}{2}\begin{pmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \hat{\boldsymbol{r}}-\boldsymbol{r}\end{pmatrix}\begin{pmatrix}\boldsymbol{X}^T\\ \boldsymbol{Z}^T\end{pmatrix}\begin{pmatrix}\boldsymbol{X} & \boldsymbol{Z}\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \hat{\boldsymbol{r}}-\boldsymbol{r}\end{pmatrix}\\
&= \text{constant} - \frac{1}{2}(\boldsymbol{y}^*-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{r})^T\boldsymbol{B}(\boldsymbol{y}^*-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{r}),\\
\ell_2 &= -\frac{1}{2}\sum_{i=1}^{m}[q_i\log(2\pi\sigma^2)+\log(|\boldsymbol{A}_i(\boldsymbol{\phi})|)+\sigma^{-2}\boldsymbol{r}_i^T\boldsymbol{A}_i^{-1}(\boldsymbol{\phi})\boldsymbol{r}_i],
\end{aligned}$$

and $\boldsymbol{y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{r}}$ where q_i indicates the dimension of $\boldsymbol{r}_i$. To follow the procedure that REML estimators of variance components are derived from BLUP estimators $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{r}}$ in a normal error model, the response vector $\boldsymbol{y}$, which is not necessarily normally distributed in the GLMM, can be replaced with $\boldsymbol{y}^*$. In addition, replace $\boldsymbol{D}^{-1}$ with $\boldsymbol{B}$ and set $\sigma^2 = 1$ to obtain approximate REMQL estimators ($\hat{\boldsymbol{\phi}}_{\text{REMQL}}$) of variance components. In the manner of Patterson and Thompson (1971)[34], the loglikelihood ℓ^* can be rewritten by

$$\ell^*_{\text{REMQL}} = -\frac{1}{2}[(n-p)\log 2\pi + \log|\boldsymbol{K}\boldsymbol{\Sigma}\boldsymbol{K}| + \boldsymbol{y}^{*T}\boldsymbol{K}(\boldsymbol{K}\boldsymbol{\Sigma}\boldsymbol{K})^{-1}\boldsymbol{K}\boldsymbol{y}^*]$$

where $\boldsymbol{\Sigma} = \boldsymbol{B}^{-1} + \boldsymbol{Z}\boldsymbol{A}\boldsymbol{Z}^T$, $\boldsymbol{K} = \boldsymbol{B} - \boldsymbol{B}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{B}$. Finally, we can obtain $\hat{\boldsymbol{\phi}}_{\text{REMQL}}$ by solving the first order derivative of ℓ^*_{REMQL} with respect to $\boldsymbol{\phi}$, which is

$$\frac{\partial \ell^*_{\text{REMQL}}}{\partial \boldsymbol{\phi}} = \frac{1}{2}\left[\text{tr}\left(\boldsymbol{A}(\boldsymbol{\phi})^{-1}\frac{\partial \boldsymbol{A}(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right) + \text{tr}\left(\boldsymbol{V}_{\boldsymbol{r}}^*\frac{\partial \boldsymbol{A}(\boldsymbol{\phi})^{-1}}{\partial \boldsymbol{\phi}}\right) + \tilde{\boldsymbol{r}}^T\tilde{\boldsymbol{r}}\frac{\partial \boldsymbol{A}(\boldsymbol{\phi})^{-1}}{\partial \boldsymbol{\phi}}\right] = 0 \quad (2.12)$$

where $\boldsymbol{V}_{\boldsymbol{r}}^*$ indicates the block matrix portion of the inverse of $\boldsymbol{V}$ corresponding to $\boldsymbol{r}$, and

$$\boldsymbol{V} = \begin{pmatrix}\boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{B}\boldsymbol{Z}\\ \boldsymbol{Z}^T\boldsymbol{B}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{B}\boldsymbol{Z}+\boldsymbol{A}^{-1}\end{pmatrix}$$

which is analogous to the original matrix in (2.11) with $\boldsymbol{B}$ in place of $\boldsymbol{D}^{-1}$ (see McGilchrist, 1994[25]; MaGilchrist and Yau, 1995[26] for the detail). In summary, the BLUP (REMQL) estimation procedure for the GLMM is as follows:

Step 1. Establish the BLUP-type loglikelihood of $\boldsymbol{y}$ and $\boldsymbol{r}$, using an $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{r}$.

Step 2. Estimate $\boldsymbol{\beta}$ and $\boldsymbol{r}$, using the Newton-Raphson (N-R) iterative procedure with initial estimates of $\boldsymbol{\beta}$, $\boldsymbol{r}$, and $\boldsymbol{\phi}$.

Step 3. Find the REMQL estimators ($\hat{\boldsymbol{\phi}}_{REMQL}$) of variance components, using REMQL estimating equations (2.12) with the BLUP estimators $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{r}}$ from step 2.

Step 4. At each iteration, update initial estimates with the estimates in the previous iteration until convergence.

3.0 PROPOSED BLUP (REMQL) ESTIMATION IN THE POISSON HURDLE MODEL WITH CORRELATED RANDOM EFFECTS

3.1 NOTATION AND MODEL SPECIFICATION

Following the notation in Section 2.5, the Poisson hurdle (H) model in (2.6) with random effects by Min and Agresti (2005)[29] is used. In the regression setting, both logit($p_{H,ij}$) and log(μ_{ij}) are assumed to depend on linear functions of the covariates in (2.7). With the correlation between the binomial and count components in the covariance matrix $\boldsymbol{A}_i(\boldsymbol{\phi})$ in (2.8), the estimation is done jointly in the correlated random effects.

We adapt the framework of McGilchrist (1994)[25] and McGilchrist and Yau (1995)[26] to develop the BLUP (REMQL) estimation of the Poisson H model in Section 2.5 with random effects and correlated components. The BLUP-type loglikelihood (joint loglikelihood) of Y_{ij} and $\boldsymbol{r}_i$ in (2.10) can be rewritten as $\ell(\boldsymbol{y},\boldsymbol{r}) = \ell_1(\boldsymbol{y}|\boldsymbol{r}) + \ell_2(\boldsymbol{r})$, where

$$\begin{aligned} \ell_1(\boldsymbol{y}|\boldsymbol{r}) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}[\mathrm{I}(y_{ij}=0)\log p_{H,ij} + (1-\mathrm{I}(y_{ij}=0))\log(1-p_{H,ij}) \\ &+ (1-\mathrm{I}(y_{ij}=0))\left\{-\mu_{ij} + y_{ij}\log\mu_{ij} - \log(y_{ij}!) - \log(1-e^{-\mu_{ij}})\right\}], \\ \ell_2(\boldsymbol{r}) &= \text{constant} - \frac{1}{2}\sum_{i=1}^{m}\left[\log(|\boldsymbol{A}_i(\boldsymbol{\phi})|) + \boldsymbol{r}_i^T\boldsymbol{A}_i(\boldsymbol{\phi})^{-1}\boldsymbol{r}_i\right], \end{aligned} \tag{3.1}$$

I($\cdot$) represents a binary indicator function, $\boldsymbol{y}$ denotes a vector of y_{ij}, and $\boldsymbol{r} = (\boldsymbol{r}_1^T, \boldsymbol{r}_2^T, ..., \boldsymbol{r}_m^T)$. Here, $\ell_1(\boldsymbol{y}|\boldsymbol{r})$ is the loglikelihood function when the random effects are conditionally fixed, and $\ell_2(\boldsymbol{r})$ indicates the penalty function for the conditional loglikelihood. First, coefficients $(\boldsymbol{\alpha},\boldsymbol{\beta})$ in the linear predictors are estimated for fixed variance components by maximizing the above BLUP-type loglikelihood. Then, we can estimate the variance component parameters $\boldsymbol{\phi} = (\sigma_u, \sigma_v, \rho)$ by using REMQL estimating equations in (2.12). Estimation can be done

iteratively via the N-R algorithm. Suppose $\boldsymbol{\xi} = \boldsymbol{W\alpha} + \boldsymbol{Ru}$ and $\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Rv}$ where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{r}^T)^T$ is the vector of unknown parameters of interest, and $\boldsymbol{R}$ is a design matrix for random components. In the initial step, coefficients in the linear predictors $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{r})$ are estimated given initial values $\boldsymbol{\theta}_0$ by

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \boldsymbol{V}^{-1}\frac{\partial \ell}{\partial \boldsymbol{\theta}}, \qquad \boldsymbol{V} = -\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}, \tag{3.2}$$

where $\boldsymbol{V}$ indicates the negative second derivatives of the BLUP-type loglikelihood (ℓ) with respect to $\boldsymbol{\theta}$. From the BLUP-type loglikelihood, we can obtain:

$$\begin{aligned}
\frac{\partial \ell}{\partial \boldsymbol{\alpha}} &= \boldsymbol{W}^T\frac{\partial \ell_1}{\partial \boldsymbol{\xi}}, \qquad \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T\frac{\partial \ell_1}{\partial \boldsymbol{\eta}}, \qquad \frac{\partial \ell_1}{\partial \boldsymbol{u}} = \boldsymbol{R}^T\frac{\partial \ell_1}{\partial \boldsymbol{\xi}}, \qquad \frac{\partial \ell_1}{\partial \boldsymbol{v}} = \boldsymbol{R}^T\frac{\partial \ell_1}{\partial \boldsymbol{\eta}}, \qquad (3.3)\\
\frac{\partial \ell}{\partial \boldsymbol{r}} &= \begin{pmatrix} \frac{\partial \ell_1}{\partial \boldsymbol{u}} \\ \frac{\partial \ell_1}{\partial \boldsymbol{v}} \end{pmatrix} \boldsymbol{G} - \boldsymbol{A}^{-1}\boldsymbol{r},
\end{aligned}$$

where a simple $2m \times 2m$ matrix (G) satisfies $\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{v} \end{pmatrix} G = \boldsymbol{r}$ and $\boldsymbol{A} = [\boldsymbol{A}_1(\boldsymbol{\phi}), \boldsymbol{A}_2(\boldsymbol{\phi}), ..., \boldsymbol{A}_m(\boldsymbol{\phi})]$ denotes a block diagonal matrix. Now,

$$\begin{aligned}
\frac{\partial \ell_1}{\partial \xi_{ij}} &= \mathrm{I}(y_{ij} = 0) - \frac{e^{\xi_{ij}}}{1 + e^{\xi_{ij}}},\\
\frac{\partial \ell_1}{\partial \eta_{ij}} &= (1 - \mathrm{I}(y_{ij} = 0)) \left\{ y_{ij} - \frac{e^{\eta_{ij}}}{1 - exp(-e^{\eta_{ij}})} \right\}.
\end{aligned}$$

The second derivatives of the BLUP-type loglikelihood are obtained as follows:

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}} &= \boldsymbol{W}^T\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi}\partial \boldsymbol{\xi}^T}\boldsymbol{W}, \qquad \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = \boldsymbol{X}^T\frac{\partial^2 \ell_1}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}}\boldsymbol{X}, \qquad \frac{\partial^2 \ell}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\beta}^T} = \boldsymbol{W}^T\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi}\partial \boldsymbol{\eta}^T}\boldsymbol{X}, \qquad (3.4)\\
\frac{\partial^2 \ell}{\partial \boldsymbol{\alpha}\partial \boldsymbol{r}^T} &= \left(\boldsymbol{W}^T\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi}\partial \boldsymbol{\xi}^T}\boldsymbol{R}\boldsymbol{W}^T\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi}\partial \boldsymbol{\eta}^T}\boldsymbol{R} \right)\boldsymbol{G}, \qquad \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\partial \boldsymbol{r}^T} = \left(\boldsymbol{X}^T\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi}\partial \boldsymbol{\eta}^T}\boldsymbol{R}\boldsymbol{X}^T\frac{\partial^2 \ell_1}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T}\boldsymbol{R} \right)\boldsymbol{G},\\
\frac{\partial^2 \ell}{\partial \boldsymbol{r}\partial \boldsymbol{r}^T} &= \boldsymbol{G}^T \begin{pmatrix} \boldsymbol{R}^T\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi}\partial \boldsymbol{\xi}^T}\boldsymbol{R} & \boldsymbol{R}^T\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi}\partial \boldsymbol{\eta}^T}\boldsymbol{R}, \\ \boldsymbol{R}^T\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi}\partial \boldsymbol{\eta}^T}\boldsymbol{R} & \boldsymbol{R}^T\frac{\partial^2 \ell_1}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T}\boldsymbol{R} \end{pmatrix}\boldsymbol{G} - \boldsymbol{A}^{-1},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi}\partial \boldsymbol{\xi}^T} &= \mathrm{Diag}\left[-\frac{e^{\boldsymbol{\xi}}}{(1 + e^{\boldsymbol{\xi}})^2} \right],\\
\frac{\partial^2 \ell_1}{\partial \boldsymbol{\eta}\partial \boldsymbol{\eta}^T} &= \mathrm{Diag}\left[-(1 - \mathrm{I}(\boldsymbol{y} = 0))\frac{e^{\boldsymbol{\eta}}(1 - exp(-e^{\boldsymbol{\eta}}) - e^{\boldsymbol{\eta}}exp(-e^{\boldsymbol{\eta}}))}{(1 - exp(-e^{\boldsymbol{\eta}}))^2} \right],\\
\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi}\partial \boldsymbol{\eta}^T} &= 0.
\end{aligned}$$

The inverse of the matrix of negative second derivatives of the BLUP-type loglikelihood with respect to $\boldsymbol{\theta}$, $\boldsymbol{V}^{-1}$, can be written as

$$\begin{bmatrix} \boldsymbol{V}^*_{\boldsymbol{\alpha}} & & \\ & \boldsymbol{V}^*_{\boldsymbol{\beta}} & \\ & & \boldsymbol{V}^*_{\boldsymbol{r}} \end{bmatrix}.$$

Asymptotic variances of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are obtained from the corresponding components $\boldsymbol{V}^*_{\boldsymbol{\alpha}}$ and $\boldsymbol{V}^*_{\boldsymbol{\beta}}$ of $\boldsymbol{V}^{-1}$.

3.2 VARIANCE COMPONENT ESTIMATION

When the N-R algorithm was performed to estimate linear predictors in the previous Section 3.1, we assumed that the variance components were known. Actually, they need to be estimated and updated in each iteration of the N-R procedure. We can obtain the approximate REMQL estimators ($\hat{\boldsymbol{\phi}}_{REMQL}$) of variance components by solving the equation (2.12) as follows:

$$tr\left(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial \boldsymbol{\phi}}\right) + tr\left(\boldsymbol{V}^*_{\boldsymbol{r}}\frac{\partial \boldsymbol{A}^{-1}}{\partial \boldsymbol{\phi}}\right) + \boldsymbol{r}^T\boldsymbol{r}\frac{\partial \boldsymbol{A}^{-1}}{\partial \boldsymbol{\phi}} = 0. \tag{3.5}$$

Note that

$$\frac{\partial \boldsymbol{A}_i}{\partial \sigma_u} = \begin{pmatrix} 2\sigma_u & \rho\sigma_v \\ \rho\sigma_v & 0 \end{pmatrix}, \qquad \frac{\partial \boldsymbol{A}_i}{\partial \sigma_v} = \begin{pmatrix} 0 & \rho\sigma_u \\ \rho\sigma_u & 2\sigma_v \end{pmatrix}, \qquad \frac{\partial \boldsymbol{A}_i}{\partial \rho} = \begin{pmatrix} 0 & \sigma_u\sigma_v \\ \sigma_u\sigma_v & 0, \end{pmatrix}, \tag{3.6}$$

and

$$\begin{aligned} \frac{\partial \boldsymbol{A}_i^{-1}(\boldsymbol{\phi})}{\partial \sigma_u} &= \frac{1}{\sigma_u^3\sigma_v^2(1-\rho^2)}\begin{pmatrix} -2\sigma_v^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & 0 \end{pmatrix}, \\ \frac{\partial \boldsymbol{A}_i^{-1}(\boldsymbol{\phi})}{\partial \sigma_v} &= \frac{1}{\sigma_u^2\sigma_v^3(1-\rho^2)}\begin{pmatrix} 0 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & -2\sigma_u^2 \end{pmatrix}, \\ \frac{\partial \boldsymbol{A}_i^{-1}(\boldsymbol{\phi})}{\partial \rho} &= \frac{1}{\sigma_u^2\sigma_v^2(1-\rho^2)^2}\begin{pmatrix} 2\rho\sigma_v^2 & -(1+\rho^2)\sigma_u\sigma_v \\ -(1+\rho^2)\sigma_u\sigma_v & 2\rho\sigma_u^2 \end{pmatrix}, \qquad i = 1, 2, ..., m. \end{aligned} \tag{3.7}$$

After substituting with (3.6) and (3.7) in (3.5), we can obtain the exact equations for the variance components $(\sigma_u, \sigma_v, \rho)$ as follows:

$$\begin{aligned}
\textstyle\sum_{i=1}^{m} \quad & \left[2\sigma_u^2\sigma_v^2(1-\rho^2) - 2\sigma_v^2 v_{ii,11} + \sigma_u\sigma_v\rho(v_{ii,12} + v_{ii,21} + 2\boldsymbol{u}_i^T\boldsymbol{v}_i) - 2\sigma_u^2\boldsymbol{u}_i^T\boldsymbol{u}_i\right] = 0, \\
\textstyle\sum_{i=1}^{m} \quad & \left[-2\sigma_u^2\sigma_v^2(1-\rho^2) - 2\sigma_u^2(v_{ii,22} + \boldsymbol{v}_i^T\boldsymbol{v}_i) + \sigma_u\sigma_v\rho(v_{ii,12} + v_{ii,21} + 2\boldsymbol{u}_i^T\boldsymbol{v}_i)\right] = 0, \\
\textstyle\sum_{i=1}^{m} \quad & [-2\sigma_u^2\sigma_v^2\rho(1-\rho^2) + 2\rho\sigma_v^2 v_{ii,11} + 2\sigma_u^2(v_{ii,22} + \boldsymbol{u}_i^T\boldsymbol{u}_i + \boldsymbol{v}_i^T\boldsymbol{v}_i) \\
& -\sigma_u\sigma_v(1+\rho^2)(v_{ii,12} + v_{ii,21} + 2\boldsymbol{u}_i^T\boldsymbol{v}_i)] = 0,
\end{aligned}$$

where v_{ii} denotes the 2×2 block matrix portion of $\boldsymbol{V}_{\boldsymbol{r}}^*$ corresponding to $\boldsymbol{r}_i$ and is also represented by $\begin{pmatrix} v_{ii,11} & v_{ii,12} \\ v_{ii,21} & v_{ii,22} \end{pmatrix}$. Finally, the N-R algorithm can estimate the variance components.

3.3 APPLICATION

We apply the correlated random effects Poisson H model to the 32-site EDCAP study to examine whether the distribution of bed days varies by intervention arm and PSI risk class among low risk patients. To achieve our study aim, we only include low risk patients (N=1,877) in CT and PA who have clinical and radiographic evidence of pneumonia and a PSI risk class$\leq$3 without hypoxemia. There are three guideline implementation interventions: low intensity (8 sites); moderate intensity (12 sites); and high intensity (12 sites). Among eligible low risk patients, 57% (N=1,061) of the patients were treated as outpatients, and 43% (N=816) were treated as inpatients. The bed days of those patients are the outcome variable of interest; we consider the patient-level PSI risk class and the site-level intervention arm as potential risk factors. We included PSI risk class and intervention arm as dummy variables in the model: PSI2 (1 if PSI=2; 0 else); PSI3 (1 if PSI=3; 0 else); Mod (1 if moderate intensity intervention; 0 else); High (1 if high intensity intervention; 0 else). The Poisson H model with random effects can be written as:

$$\begin{aligned}
\text{logit}(p_{H,ij}) &= \alpha_0 + \alpha_1 \cdot \text{PSI2} + \alpha_2 \cdot \text{PSI3} + \alpha_3 \cdot \text{Mod} + \alpha_4 \cdot \text{High} + \boldsymbol{u}_i, \\
\log(\mu_{ij}) &= \beta_0 + \beta_1 \cdot \text{PSI2} + \beta_2 \cdot \text{PSI3} + \beta_3 \cdot \text{Mod} + \beta_4 \cdot \text{High} + \boldsymbol{v}_i,
\end{aligned}$$

where $(\boldsymbol{u}_i, \boldsymbol{v}_i)^T$ is assumed to be distributed as $N(\mathbf{0}, \boldsymbol{A}_i(\phi))$ when the covariance matrix $\boldsymbol{A}_i(\phi)$ is defined as before with i=1, ..., 32.

Table 1 summarizes the eligible low risk patients in the EDCAP study. Among low risk patients, thirty-seven percent have PSI=1, 37% have PSI=2, and 26% have PSI=3. Lower PSI risk class is associated with higher probability of outpatient care (0.82 for PSI=1; 0.51 for PSI=2; 0.28 for PSI=3) is with shorter average LOS of inpatients (4.0 for PSI=1; 4.6 for PSI=2; 5.8 for PSI=3). Twenty-three percent of the patients presented to low intensity intervention sites, 40% presented to moderate intensity intervention sites, and 37% presented to high intensity intervention sites. Both moderate and high intensity intervention sites have a higher proportion of low risk outpatients than the low intensity intervention sites (moderate, 0.62; high, 0.63 vs. low, 0.38), and the average LOS for inpatients differs slightly by intervention arm. Overall average bed days is highest for the low intensity intervention sites.

For example, zero-truncated Poisson or zero-truncated negative binomial regression can account only for inpatient LOS, and Poisson or negative binomial regression do not accurately predict short bed days in the EDCAP data (Figure 3). When we used the H models for bed days with the EDCAP data (Figure 4), we can see that both the Poisson H and negative binomial H models adequately account for excess zeros, although the Poisson H does not fit the non-zero part as well as the negative binomial H does.

Table 2 presents the results using the Poisson H model with (a) uncorrelated random effects ($\rho = 0$) based on BLUP (REMQL) estimation, (b) correlated random effects ($\rho \neq 0$) based on BLUP (REMQL) estimation, and (c) correlated random effects based on ML estimation. First, when we compare (a) with (b), the results are almost identical between the two models, regardless of the bivariate correlation, except for a small difference in AIC. When we look at the results between (b) and (c), the fixed effects estimates and standard errors are almost identical between the two estimation methods for both components of the model. However, the AIC is somewhat smaller for the BLUP (REMQL) model, indicating a relatively better fit. The estimated bivariate correlation between the two components is low (-0.04 for BLUP (REMQL); -0.01 for ML). Based on (b), the log odds ratio (log OR) of outpatient care decreases significantly with increasing risk class, with log ORs of -1.47 and

-2.51 for PSI2 and PSI3, respectively, and increases significantly at the moderate and high intensity intervention sites, with log ORs of 0.97 and 0.91 for moderate and high intensity sites, respectively (Table 2).

Finally, Figure 5 represents the random effect predictions ($\hat{\boldsymbol{u}}_i$ and $\hat{\boldsymbol{v}}_i$) in (2.7) for the logistic and the Poisson parts, respectively, when we used the correlated random effects Poisson H model based on the BLUP (REMQL) estimation. We observed that site 25 is a slightly unusual in that the predicted random effect for the logistic part is relatively small while the predicted random effect for the Poisson part is close to 0; site 25 has relatively low proportion of outpatient care for a moderate intensity site. Figure 5 indicates that there is more site-level variation in the logistic part (hospitalization decision) than in the Poisson part.

3.4 SIMULATION STUDY

We conducted simulation studies to compare the performance of the proposed BLUP (REMQL) to ML in the correlated random effects Poisson H model with a plausible range of bivariate correlations. We designed the simulation to mimic the structure of the EDCAP data. We considered an unbalanced cluster-randomized study, including patient-level covariates (PSI1, PSI2, PSI3) and site-level covariates (Low, Mod, High). PSI1 and low intensity intervention are the reference levels. From each of the $m = 32$ sites, n_i patients are randomly generated from the Poisson distribution. Based upon the preliminary analysis of the EDCAP data (Table 2), $\boldsymbol{\alpha}$ is chosen as (0.8, -1.5, -2.5, 1.0, 0.9), $\boldsymbol{\beta}$ is chosen as (1.4, 0.1, 0.3, -0.1, 0.1), $\sigma_u = 0.6$, $\sigma_v = 0.2$, and ρ takes one of the following values (-0.1, -0.3, -0.5, -0.7). The number of replications is 1,000 for each of the four simulated settings.

To verify whether our simulated data are similar to the EDCAP data, we presented the proportion for each PSI risk class (1, 2, 3), the proportion for each intervention intensity (low, moderate, high), and the proportion of outpatient by each subgroup (Table 3). Most of the summary values in Table 3 are very similar to those in Table 1. We conclude that the bed days of the simulated data reasonably represent the bed days of the EDCAP data,

because the cumulative distribution of bed days by both are almost identical (Figure 6).

Tables 4 and 5, which summarize the results of the simulation studies, report the average bias, the relative bias to the true parameter (Percent), standard error (SE), mean square error (MSE), and coverage probability (CP) of the 95% confidence interval over 1,000 replications for each value of ρ. The relative bias is calculated by 100×abs (Bias/True). For fixed effect estimates in the Poisson part, both the ML and BLUP (REMQL) give negligible biases relative to their corresponding SEs as well as small MSEs, while the BLUP (REMQL) fixed effects estimates in the logistic part have slightly larger but still small biases than the ML estimates (all percents $\leq$2.0). For the variance components $(\sigma_u, \sigma_v, \rho)$ of random effects, the BLUP (REMQL) estimates have much smaller biases than the ML estimates (the smallest ratios: 7.3% vs. 2.3% for σ_u (Table 5(c)); 7.5% vs. 2.0% for σ_v (Table 4(a)); 63.7% vs. 1.7% for ρ (Table 4(b))). The proposed BLUP (REMQL) approach appears to have a slightly better coverage probability across the scenarios considered. For the simulated dataset generated with highly correlated random effects ($\rho = -0.7$), the BLUP (REMQL) estimate of the bivariate correlation has a much smaller bias than the ML estimate (0.002 for BLUP (REMQL); 0.455 for ML).

In summary, the simulation results demonstrate that the proposed estimation in the Poisson H model with correlated random effects performs well for the linear predictors and variance components considered. We used the same convergence criteria for both estimation methods; all replications converged for the BLUP (REMQL), while some replications did not converge for the ML (1/1000 replications at $\rho = -0.1$; 3/1000 replications at $\rho = -0.3$; 2/1000 replications at $\rho = -0.5$; 9/1000 replications at $\rho = -0.7$). The BLUP (REMQL) runs in about 1/7 the time as ML. We used the SAS procedure NLMIXED to fit this model with ML (Appendix C), and we used R to obtain the proposed BLUP (REMQL) estimates (Appendix A).

We also can conclude that the predicted random effects based on the BLUP (REMQL) estimation reflect the original bivariate correlation, which was plotted using the simulated data (Figure 7).

3.5 DISCUSSION

In this chapter, we proposed a BLUP (REMQL) approach to estimate the parameters of a Poisson H model with correlated random effects. We also illustrated the application of the Poisson H model to a potential efficiency metric, bed days, in health services studies. This model appropriately account for excess zeros, clustering by site, and a possible bivariate correlation between the binary and count components of this model. This model gives an overall assessment of the intervention effect on two aspects of care, e.g., admission and LOS in the EDCAP study. While the interventions in EDCAP were designed to influence the admission decision and increase performance of recommended processes of care in the Emergency Department (ED), no intervention influenced LOS for inpatients. Our results confirmed that the higher intensity interventions significantly reduced hospitalizations but did not affect LOS. In addition, the low negative bivariate correlation indicates that sites with relatively low admission rates for low risk patients tended to have somewhat shorter LOS for those admitted. We also identified the unusual sites (hospitals) and used prediction site effects to investigate associations between admission and LOS across sites. This information could contribute to identifying efficient sites with good intervention performance, particularly in the case where the intervention addresses both outcomes.

As McCulloch (2001)[24] pointed out, the BLUP (REMQL) can produce biased estimates with unbalanced binary data, although this approach has been shown to reduce the bias in variance components relative to ML estimation in a GLMM setting. Our simulation study confirmed that the BLUP (REMQL) approach provides less biased estimates of variance components than ML in a random effects Poisson H model, and gives estimates similar to the ML for the linear predictors in the Poisson part. The BLUP (REMQL) estimates of the linear predictors in the logistic part appeared to be somewhat more biased than the ML estimates in our unbalanced study. However, the proposed BLUP (REMQL) ran considerably faster than the ML and demonstrated better convergence properties.

The proposed Poisson H model with correlated random effects has accounted for overdispersion from excess zeros and possible clustering by site. However, we might be required to consider overdispersion drive to heterogeneity across sites. In Chapter 4, we will present in

detail BLUP (REMQL) estimation in the negative binomial H model with correlated random effects.

Table 1: Probability of outpatient, mean and median of inpatients bed days, and mean and median of overall bed days by PSI risk class and by intervention arm for 1877 eligible low risk patients

			Bed Days	
			Inpatients	**Overall**
	n	**Pr(Outpatient)**	**Mean(Median)**	**Mean(Median)**
PSI risk class				
1	697	0.82	4.0 (3.0)	0.7 (0.0)
2	691	0.51	4.6 (4.0)	2.2 (0.0)
3	486	0.28	5.8 (4.0)	4.1 (3.0)
Intervention				
Low	438	0.38	5.0 (4.0)	3.1 (2.0)
Mod	748	0.62	4.9 (4.0)	1.9 (0.0)
High	691	0.63	5.1 (4.0)	1.9 (0.0)
Overall		0.57	5.0 (4.0)	2.2 (0.0)

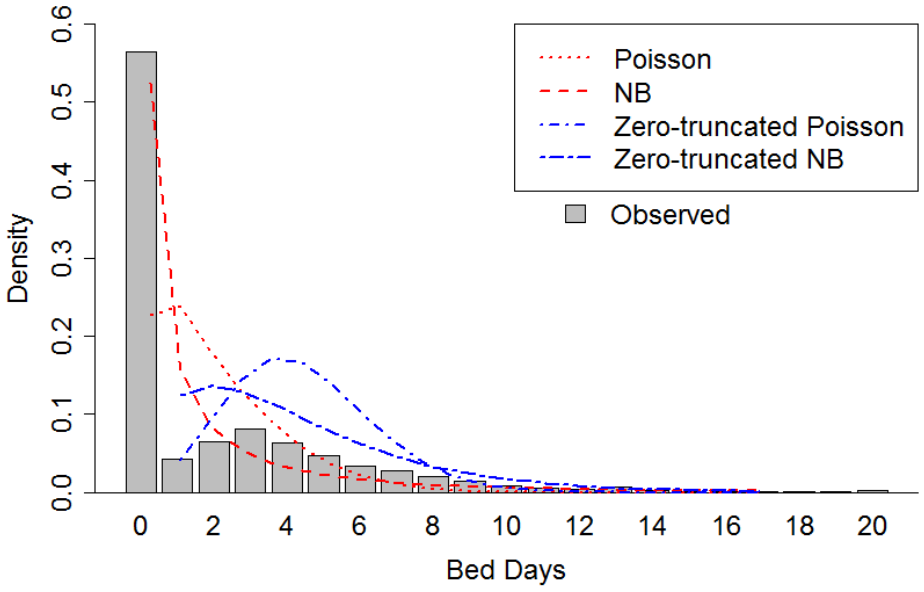

Figure 3: Observed vs predicted distribution of bed days by simple models (Poisson, NB, Zero-truncated Poisson, Zero-truncated NB)

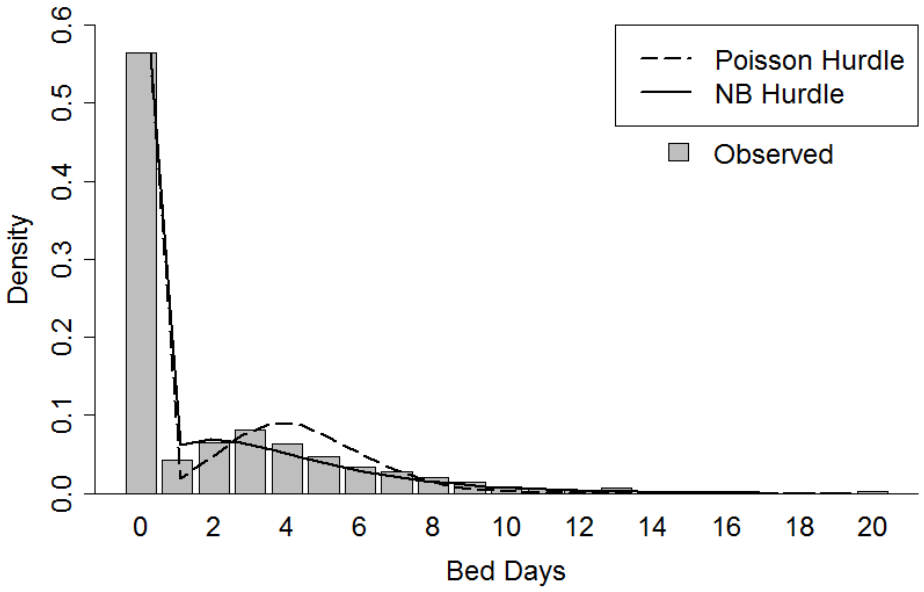

Figure 4: Observed vs predicted distribution of bed days by hurdle models

Table 2: Poisson H model estimates with (a) uncorrelated random effects based on BLUP (REMQL) estimation, (b) correlated random effects based on BLUP (REMQL) estimation, and (c) correlated random effects based on ML estimation

	(a) $\boldsymbol{\rho} = 0$			(b) $\boldsymbol{\rho} \neq 0$			(c) $\boldsymbol{\rho} \neq 0$		
	BLUP (REMQL)			**BLUP (REMQL)**			**ML**		
	Estimate	**SE**	**P-value**	**Estimate**	**SE**	**P-value**	**Estimate**	**SE**	**P-value**
				Logistic Part: Pr(Outpatient)					
Cons	0.79	0.25	<.01	0.79	0.25	<.01	0.80	0.24	<.01
PSI2	-1.47	0.13	<.001	-1.47	0.13	<.001	-1.49	0.13	<.001
PSI3	-2.51	0.15	<.001	-2.51	0.15	<.001	-2.54	0.15	<.001
Mod	0.97	0.30	<.01	0.97	0.30	<.01	0.98	0.29	<.01
High	0.91	0.30	<.01	0.91	0.30	<.01	0.93	0.29	<.01
$\boldsymbol{\sigma}_u$	0.57			0.57			0.54		
				Poisson Part: Inpatient Bed Days					
Cons	1.41	0.09	<.001	1.41	0.09	<.001	1.41	0.08	<.001
PSI2	0.09	0.05	.09	0.09	0.05	.09	0.09	0.05	.09
PSI3	0.34	0.05	<.001	0.34	0.05	<.001	0.34	0.05	<.001
Mod	-0.06	0.10	.51	-0.06	0.10	.51	-0.06	0.09	.49
High	0.02	0.10	.88	0.02	0.10	.87	0.02	0.09	.87
$\boldsymbol{\sigma}_v$	0.19			0.19			0.18		
$\boldsymbol{\rho}$				**-0.04**			**-0.01**		
-2*ℓ		6503.6			6503.6			6647.1	
AIC		6527.6			6529.6			6673.1	

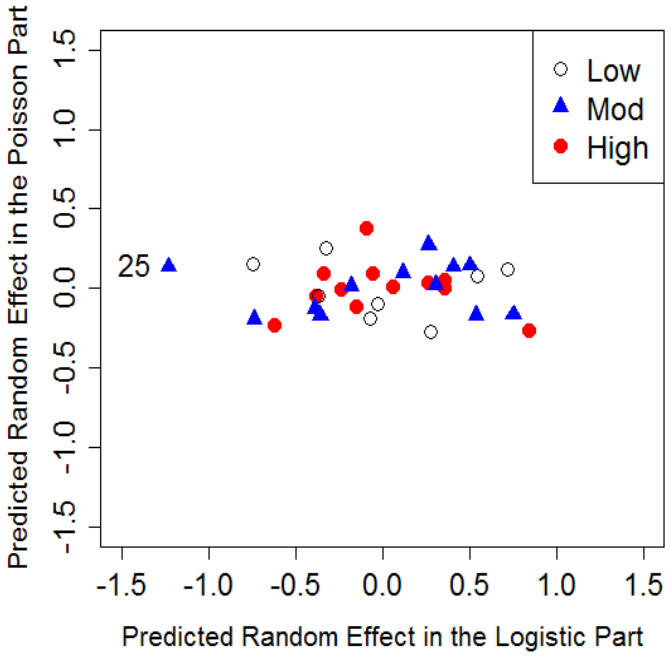

Figure 5: Site specific predicted random effects for the logistic and Poisson parts of the Poisson H model for the low (○), moderate (▲), and high (●) intensity intervention sites

Table 3: Probability of outpatient, mean and median of inpatients bed days, and mean and median of overall bed days by PSI risk class and by intervention arm for one simulated dataset (N=1,890)

			Bed Days	
	n	**Pr(Outpatient)**	**Inpatients Mean(Median)**	**Overall Mean(Median)**
PSI risk class				
1	697	0.82	5.0 (5.0)	0.9 (0.0)
2	707	0.55	5.0 (5.0)	2.2 (0.0)
3	486	0.34	6.1 (6.0)	4.1 (4.0)
Intervention				
Low	504	0.40	5.4 (5.0)	3.3 (3.0)
Mod	529	0.66	4.9 (5.0)	1.6 (0.0)
High	857	0.67	5.9 (6.0)	2.0 (0.0)
Overall		0.59	5.5 (5.0)	2.2 (0.0)

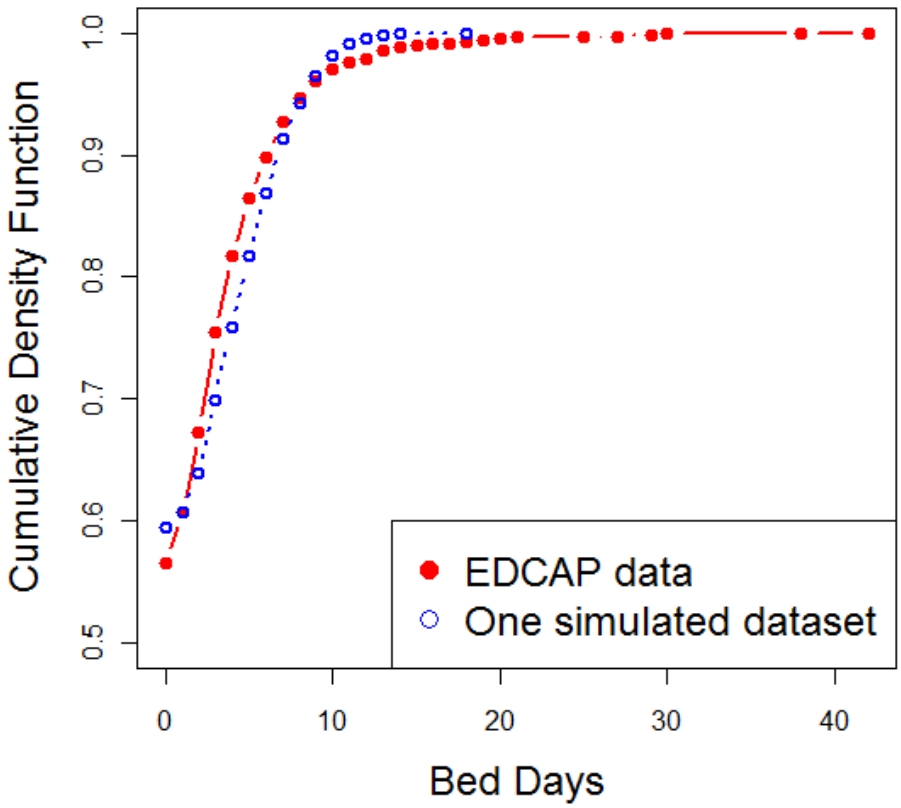

Figure 6: Cumulative density function of bed days by EDCAP data (•) and one simulated dataset (◦) with $\rho = -0.1$

Table 4: Simulation results using correlated random effects Poisson H model based on ML and BLUP (REMQL) estimation with 1000 replications and a plausible range of bivariate correlation ($\rho = -0.1, -0.3$)

		Bias(Percent*)		SE		MSE		CP	
Parameter	**True**	**ML**	**REMQL**	**ML**	**REMQL**	**ML**	**REMQL**	**ML**	**REMQL**
(a) $\rho = -0.1$				**Logistic Part: Pr(Outpatient)**					
α_0 : **Cons**	0.8	0.002(0.3)	-0.007(0.9)	0.246	0.254	0.130	0.133	0.94	0.95
α_1 : **PSI2**	-1.5	0.000(-)	0.017(1.1)	0.130	0.129	0.033	0.033	0.96	0.96
α_2 : **PSI3**	-2.5	0.008(0.3)	0.037(1.5)	0.150	0.147	0.045	0.045	0.95	0.94
α_3 : **Mod**	1.0	-0.005(0.5)	-0.016(1.6)	0.303	0.315	0.198	0.204	0.93	0.94
α_4 : **High**	0.9	-0.006(0.7)	-0.015(1.7)	0.304	0.316	0.202	0.207	0.93	0.94
σ_u	0.6	-0.040(6.7)	-0.011(1.8)						
				Poisson Part: Inpatient Bed Days					
β_0 : **Cons**	1.4	-0.007(0.5)	-0.003(0.2)	0.085	0.088	0.017	0.016	0.94	0.95
β_1 : **PSI2**	0.1	0.001(1.0)	0.001(1.0)	0.050	0.050	0.005	0.005	0.95	0.95
β_2 : **PSI3**	0.3	0.002(0.7)	0.002(0.7)	0.049	0.049	0.005	0.005	0.96	0.96
β_3 : **Mod**	-0.1	0.004(4.0)	0.005(5.0)	0.099	0.103	0.020	0.021	0.94	0.95
β_4 : **High**	0.1	0.010(10.0)	0.009(9.0)	0.098	0.103	0.022	0.021	0.93	0.94
σ_v	0.2	-0.015(7.5)	-0.004(2.0)						
ρ	-0.1	0.064(64.0)	-0.001(1.0)						
(b) $\rho = -0.3$				**Logistic Part: Pr(Outpatient)**					
α_0 : **Cons**	0.8	0.007(0.9)	-0.003(0.4)	0.243	0.252	0.125	0.128	0.93	0.94
α_1 : **PSI2**	-1.5	-0.007(0.5)	0.009(0.6)	0.130	0.129	0.035	0.034	0.94	0.94
α_2 : **PSI3**	-2.5	-0.007(0.3)	0.022(0.9)	0.150	0.147	0.044	0.043	0.95	0.95
α_3 : **Mod**	1.0	0.003(0.3)	-0.008(0.8)	0.302	0.314	0.189	0.195	0.93	0.95
α_4 : **High**	0.9	-0.008(0.9)	-0.018(2.0)	0.302	0.314	0.184	0.190	0.94	0.95
σ_u	0.6	-0.042(7.0)	-0.012(2.0)						
				Poisson Part: Inpatient Bed Days					
β_0 : **Cons**	1.4	0.000(-)	0.002(1.1)	0.085	0.088	0.015	0.016	0.93	0.94
β_1 : **PSI2**	0.1	0.001(1.0)	0.001(1.0)	0.050	0.050	0.005	0.005	0.95	0.95
β_2 : **PSI3**	0.3	0.002(0.7)	0.002(0.7)	0.049	0.049	0.005	0.005	0.96	0.96
β_3 : **Mod**	-0.1	-0.003(3.0)	-0.001(1.0)	0.099	0.104	0.020	0.021	0.93	0.95
β_4 : **High**	0.1	0.000(-)	0.000(-)	0.098	0.103	0.020	0.021	0.93	0.94
σ_v	0.2	-0.013(6.5)	-0.001(0.5)						
ρ	-0.3	0.191(63.7)	0.005(1.7)						

* indicates the relative bias to the true parameter, which is calculated by 100*abs (Bias/True).

For ML, 1 replication did not converge when $\rho = -0.1$ and 3 replications did not converge when $\rho = -0.3$.

Table 5: Simulation results using correlated random effects Poisson H model based on ML and BLUP (REMQL) estimation with 1000 replications and a plausible range of bivariate correlation ($\rho = -0.5, -0.7$)

		Bias(Percent*)		SE		MSE		CP	
Parameter	**True**	**ML**	**REMQL**	**ML**	**REMQL**	**ML**	**REMQL**	**ML**	**REMQL**
(c) $\rho = -0.5$				**Logistic Part: Pr(Outpatient)**					
α_0 : **Cons**	0.8	0.000(-)	-0.009(1.1)	0.246	0.255	0.127	0.130	0.95	0.95
α_1 : **PSI2**	-1.5	-0.008(0.5)	0.008(0.5)	0.130	0.129	0.034	0.033	0.94	0.94
α_2 : **PSI3**	-2.5	-0.004(0.2)	0.023(0.9)	0.150	0.147	0.045	0.044	0.95	0.95
α_3 : **Mod**	1.0	-0.001(0.1)	-0.011(1.1)	0.303	0.315	0.192	0.198	0.94	0.95
α_4 : **High**	0.9	0.003(0.3)	-0.006(0.7)	0.303	0.316	0.190	0.196	0.96	0.96
σ_u	0.6	-0.044(7.3)	-0.014(2.3)						
				Poisson Part: Inpatient Bed Days					
β_0 : **Cons**	1.4	0.001(0.1)	0.004(0.3)	0.085	0.088	0.016	0.016	0.94	0.95
β_1 : **PSI2**	0.1	0.003(3.0)	0.003(3.0)	0.049	0.049	0.005	0.005	0.95	0.95
β_2 : **PSI3**	0.3	0.000(-)	0.000(-)	0.048	0.048	0.005	0.005	0.93	0.93
β_3 : **Mod**	-0.1	-0.002(2.0)	0.000(-)	0.100	0.104	0.021	0.022	0.94	0.95
β_4 : **High**	0.1	-0.005(5.0)	-0.005(5.0)	0.099	0.103	0.021	0.022	0.93	0.94
σ_v	0.2	-0.014(7.0)	-0.002(1.0)						
ρ	-0.5	0.320(64.0)	0.003(0.6)						
(d) $\rho = -0.7$				**Logistic Part: Pr(Outpatient)**					
α_0 : **Cons**	0.8	0.003(0.4)	-0.005(0.6)	0.249	0.258	0.136	0.140	0.94	0.94
α_1 : **PSI2**	-1.5	-0.005(0.3)	0.008(0.5)	0.130	0.129	0.035	0.034	0.94	0.94
α_2 : **PSI3**	-2.5	-0.007(0.3)	0.016(0.6)	0.150	0.147	0.045	0.044	0.95	0.95
α_3 : **Mod**	1.0	0.003(0.3)	-0.006(0.6)	0.307	0.319	0.202	0.208	0.94	0.95
α_4 : **High**	0.9	0.003(0.3)	-0.005(0.6)	0.307	0.320	0.205	0.211	0.93	0.94
σ_u	0.6	-0.036(6.0)	-0.005(0.8)						
				Poisson Part: Inpatient Bed Days					
β_0 : **Cons**	1.4	0.000(-)	0.002(0.1)	0.085	0.088	0.016	0.016	0.93	0.94
β_1 : **PSI2**	0.1	0.002(2.0)	0.002(2.0)	0.049	0.049	0.005	0.005	0.96	0.96
β_2 : **PSI3**	0.3	0.003(1.0)	0.002(0.7)	0.048	0.048	0.005	0.005	0.95	0.95
β_3 : **Mod**	-0.1	0.001(1.0)	0.002(2.0)	0.100	0.105	0.022	0.023	0.92	0.93
β_4 : **High**	0.1	-0.003(3.0)	-0.003(3.0)	0.099	0.104	0.022	0.023	0.93	0.94
σ_v	0.2	-0.014(7.0)	-0.002(1.0)						
ρ	-0.7	0.455(65.0)	0.002(0.3)						

* indicates the relative bias to the true parameter, which is calculated by 100*abs (Bias/True).

For ML, 2 replication did not converge when $\rho = -0.5$ and 9 replications did not converge when $\rho = -0.7$.

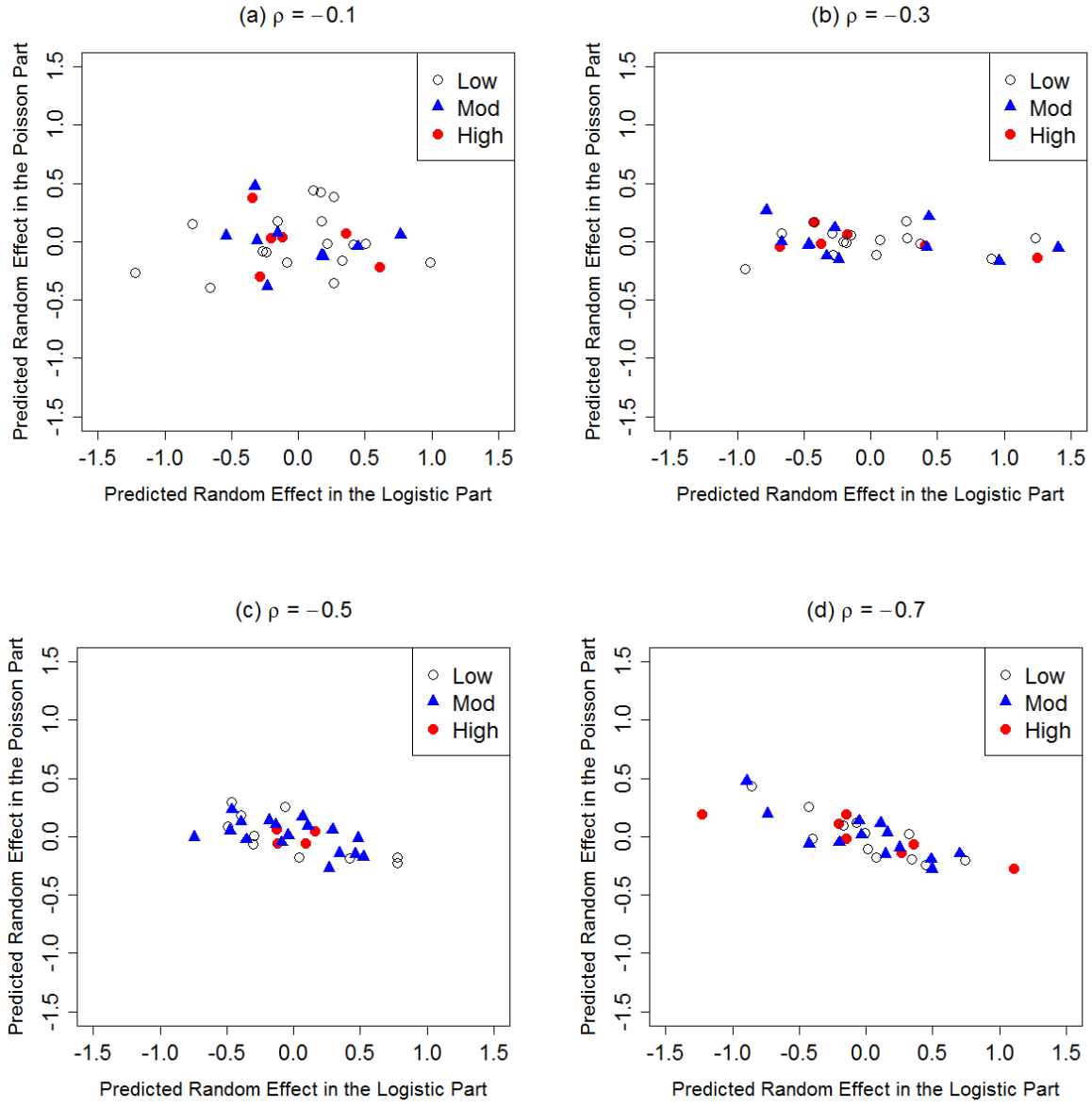

Figure 7: Random site effect predictions for simulated datasets with (a) $\rho = -0.1$, (b) $\rho = -0.3$, (c) $\rho = -0.5$, and (d) $\rho = -0.7$ for the low (○), moderate (▲), and high (●) intensity intervention sites

4.0 PROPOSED BLUP (REMQL) ESTIMATION IN THE NEGATIVE BINOMIAL HURDLE MODEL WITH CORRELATED RANDOM EFFECTS

The Poisson hurdle model with correlated random effects was proposed to account for both overdispersion from excess zeros and clustering by site. However, it does not account for overdispersion from heterogeneity between sites. This fact leads us to use the negative binomial hurdle model with correlated random effects, since the negative binomial distribution relaxes the assumption of the Poisson distribution that the mean is equal to the variance.

4.1 NOTATION AND MODEL SPECIFICATION

The negative binomial hurdle (H) model with random effects can be obtained by replacing f in (2.6) with the negative binomial distribution. Let Y_{ij} $(i = 1, 2, ..., m; j = 1, 2, ..., n_i)$ be the number of bed days of patient j at site i, where m is the number of sites, n_i is the number of patients at site i, and $n = \sum_{i=1}^{m} n_i$ is the total number of patients. Then, the negative binomial H model with random effects is:

$$\begin{aligned} P(Y_{ij} = 0) &= p_{H,ij}, \\ P(Y_{ij} = y_{ij} | y_{ij} > 0) &= (1 - p_{H,ij}) \cdot \frac{f(y_{ij})}{1 - f(0)} \\ &= (1 - p_{H,ij}) \begin{pmatrix} y_{ij} + k - 1 \\ y_{ij} \end{pmatrix} \frac{t_{ij}^k (1 - t_{ij})^{y_{ij}}}{1 - t_{ij}^k}, \end{aligned} \tag{4.1}$$

where $p_{H,ij}$ indicates the conditional probability of not being hospitalized, given patient j at site i is at risk for hospitalization; $t_{ij} = \frac{k}{k+\mu_{ij}}$, k is the scale parameter, which is

equal to 1/dispersion parameter. Here, μ_{ij} is the mean of the underlying negative binomial distribution. Hence, $\frac{1}{k}$ indicates the extra-variation parameter (i.e., $Var(\boldsymbol{y}) = E(\boldsymbol{y}) + \frac{1}{k} \times E(\boldsymbol{y})^2$). We can regard $f(y_{ij})/(1-f(0))$ as a truncated negative binomial distribution. When k goes to infinity, the negative binomial H model reduces to the Poisson H model. We can assume that both logit$(p_{H,ij})$ and log(μ_{ij}) depend upon the linear functions of the covariates in the regression setting, similar to (2.7). In the comparable random effects Poisson H model, the BLUP-type loglikelihood of Y_{ij} and r_i can be written as $\ell(\boldsymbol{y}, \boldsymbol{r}) = \ell_1(\boldsymbol{y}|\boldsymbol{r}) + \ell_2(\boldsymbol{r})$, where

$$
\begin{aligned}
\ell_1(\boldsymbol{y}|\boldsymbol{r}) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}[\mathrm{I}(y_{ij}=0)\log p_{H,ij} + (1-\mathrm{I}(y_{ij}=0))\log(1-p_{H,ij}) \qquad (4.2)\\
&+ (1-\mathrm{I}(y_{ij}=0))\left\{\log\frac{\Gamma(y_{ij}+k)}{\Gamma(y_{ij}+1)\Gamma(k)} + k\log t_{ij} + y_{ij}\log(1-t_{ij}) - \log(1-t_{ij}^k)\right\}],\\
\ell_2(\boldsymbol{r}) &= \text{constant} - \frac{1}{2}\sum_{i=1}^{m}\left[\log(|\boldsymbol{A}_i(\boldsymbol{\phi})|) + \boldsymbol{r}_i^T\boldsymbol{A}_i(\boldsymbol{\phi})^{-1}\boldsymbol{r}_i\right],
\end{aligned}
$$

and I$(\cdot)$ represents a binary indicator function, $\boldsymbol{y}$ is a vector of y_{ij}, and $\boldsymbol{r} = (\boldsymbol{r}_1^T, \boldsymbol{r}_2^T, ..., \boldsymbol{r}_m^T)$. First, the coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in the linear predictors are estimated for fixed variance components and scale parameter in the negative binomial distribution by maximizing the above BLUP-type loglikelihood. Then, using restricted maximum quasi-likelihood (REMQL), we can estimate the variance component parameters $\boldsymbol{\phi} = (\sigma_u, \sigma_v, \rho)$. The scale parameter k, which is assumed to be fixed in estimation of the regression coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, is also updated by maximizing a profile loglikelihood with the current estimates. Estimation can be done iteratively via the N-R algorithm. Similar to the random effects Poisson H model, we can estimate coefficients in the linear predictors $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{r}^T)^T$ given initial values $\boldsymbol{\theta}_0$ by (3.2). We implement the same estimation process with the random effects Poisson H model,

updating some derivatives as follows:

$$\begin{aligned}
\frac{\partial \ell_1}{\partial \xi_{ij}} &= \mathrm{I}(y_{ij}=0) - \frac{e^{\xi_{ij}}}{1+e^{\xi_{ij}}}, \\
\frac{\partial \ell_1}{\partial \eta_{ij}} &= (1-\mathrm{I}(y_{ij}=0)) \left\{ y_{ij} \frac{k}{k+e^{\eta_{ij}}} - \frac{k(1-\frac{k}{k+e^{\eta_{ij}}})}{1-(\frac{k}{k+e^{\eta_{ij}}})^k} \right\}, \\
\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T} &= \mathrm{Diag}\left[-\frac{e^{\xi}}{(1+e^{\xi})^2} \right], \\
\frac{\partial^2 \ell_1}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}^T} &= \mathrm{Diag}\left[-(1-\mathrm{I}(\boldsymbol{y}=0)) t(1-t) \left\{ \boldsymbol{y} - \frac{k^2(1-t)t^{k-1} - k(1-t^k)}{(1-t^k)^2} \right\} \right], \\
\frac{\partial^2 \ell_1}{\partial \boldsymbol{\xi} \partial \boldsymbol{\eta}^T} &= 0.
\end{aligned}$$

Like the random effects Poisson H model, the asymptotic variances of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ can be obtained from the corresponding components of the inverse of the matrix of negative second derivatives of the BLUP-type loglikelihood with respect to $\boldsymbol{\theta}$.

4.2 VARIANCE COMPONENT ESTIMATION

In the previous N-R algorithm to estimate linear predictors, we presumed that the parameters of the variance components are given. However, because they are not given, we need to estimate and update in each iteration of the N-R procedure to obtain the approximate REMQL estimators ($\hat{\boldsymbol{\phi}}_{REMQL}$) of the variance components by solving the REMQL estimating equations (3.5). Estimation of variance components for the random effects negative binomial H model is identical to that of the Poisson H model in Chapter 3. Hence, we do not repeat that procedure.

4.3 SCALE PARAMETER ESTIMATION

The estimation via the N-R algorithm in Section 4.1 assumed that the scale parameter k was known. In practice, k is updated and estimated in each iteration in accordance with the

updated estimates of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{r}, \sigma_u, \sigma_v$ and ρ by maximizing the profile loglikelihood function:

$$\begin{aligned} \ell_k &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}[\mathrm{I}(y_{ij}=0)\log p_{H,ij} + (1-\mathrm{I}(y_{ij}=0))\log(1-p_{H,ij}) \\ &+ (1-\mathrm{I}(y_{ij}=0))\left\{\log\frac{\Gamma(y_{ij}+k)}{\Gamma(y_{ij}+1)\Gamma(k)} + k\log t_{ij} + y_{ij}\log(1-t_{ij}) - \log(1-t_{ij}^k)\right\}]. \end{aligned} \tag{4.3}$$

Following Lee *et al.* (2003)[22], suppose $A(k) = \sum_{y_{ij}>0}\log\frac{\Gamma(y_{ij}+k)}{\Gamma(y_{ij}+1)\Gamma(k)}$, and $f(\tau) = \#\{y_{ij} \geq \tau, \forall i,j\}$ is the number of patients whose observed count is greater than or equal to τ. Then, the first and second derivatives of $A(k)$ are derived as:

$$\begin{aligned} \dot{A(k)} &= \sum_{\tau=1}^{max(y_{ij})-1}\frac{f(\tau)}{k+\tau}, \\ \ddot{A(k)} &= -\sum_{\tau=1}^{max(y_{ij})-1}\frac{f(\tau)}{(k+\tau)^2}. \end{aligned}$$

Finally, the first and second derivatives of ℓ_k can be expressed in terms of $\dot{A(k)}$ and $\ddot{A(k)}$:

$$\begin{aligned} \frac{\partial \ell_k}{\partial k} &= \dot{A(k)} + \sum_{y_{ij}>0}\frac{B_{ij}}{1-t_{ij}^k} - \frac{y_{ij}t_{ij}}{k}, \\ \frac{\partial^2 \ell_k}{\partial k^2} &= \ddot{A(k)} + \sum_{y_{ij}>0}\frac{\dot{B}_{ij}(1-t_{ij}^k) + B_{ij}^2 t_{ij}^k}{(1-t_{ij}^k)^2} + \frac{y_{ij}t_{ij}^2}{k^2}, \end{aligned} \tag{4.4}$$

where

$$B_{ij} = \log(t_{ij}) + 1 - t_{ij} \qquad \text{and} \qquad \dot{B}_{ij} = \frac{(1-t_{ij})^2}{k}.$$

The asymptotic variance of $\hat{k}$ can be obtained by $Var(\hat{k}) = (-\frac{\partial^2 \ell_k}{\partial k^2})^{-1}$.

4.4 APPLICATION

Using the same notation of in Section 3.3, the negative binomial H model with random effects can be written as:

$$\begin{aligned} \text{logit}(p_{H,ij}) &= \alpha_0 + \alpha_1 \cdot \text{PSI2} + \alpha_2 \cdot \text{PSI3} + \alpha_3 \cdot \text{Mod} + \alpha_4 \cdot \text{High} + \boldsymbol{u}_i, \\ \log(\mu_{ij}) &= \beta_0 + \beta_1 \cdot \text{PSI2} + \beta_2 \cdot \text{PSI3} + \beta_3 \cdot \text{Mod} + \beta_4 \cdot \text{High} + \boldsymbol{v}_i, \end{aligned}$$

where $(\boldsymbol{u}_i, \boldsymbol{v}_i)^T$ is assumed to be distributed as $N(\boldsymbol{0}, \boldsymbol{A}_i(\boldsymbol{\phi}))$ when the covariance matrix $\boldsymbol{A}_i(\boldsymbol{\phi})$ is defined as before with i=1,...,32.

Table 6 presents the estimates for the random effects negative binomial H model. Estimates based on ML and BLUP (REMQL) are quite similar to each other for both components, except possibly for k. Again, the AIC favors the BLUP (REMQL) model. The log odds ratios (log ORs) of -1.47 and -2.51 for PSI2 and PSI3, respectively, indicate that the log OR of outpatient care decreases significantly with increasing risk class, while the log ORs of 0.97 and 0.91 for moderate and high intensity sites, respectively, illustrate that the log OR of outpatient care increases significantly at the moderate and high intensity intervention sites (Table 6(a)). The estimated scale parameter ($k = 2.71$) suggests significant overdispersion relative to the Poisson distribution; negative binomial H model with the scale parameter accounts for extra variation, $\frac{1}{2.71} \times E(\boldsymbol{y})^2$. The estimated bivariate correlation is modest (-0.10). The p-values for the PSI2 parameters in the count component of the model are less significant in the negative binomial H model than in the Poisson H model due to the correction for overdispersion (p-value: 0.18 for negative binomial H model (Table 6(a)) vs. 0.09 for Poisson H model (Table 2(b))).

Based on the random effects negative binomial H model, the predicted site-level random effects are plotted for the logistic and the negative binomial parts (Figure 8). Again, site 25 appears to be slightly unusual but not that much different from the other sites. In addition, we found that the variation of the predicted site-level random effects in the NB part to be smaller than in the corresponding Poisson part (Figure 8 for the negative binomial H model; Figure 5 for the Poisson H model). Figure 9 illustrates the better fit of the negative binomial H model than the Poisson H model to these data.

4.5 SIMULATION STUDY

To compare the performance of the proposed BLUP (REMQL) to ML in the correlated random effects negative binomial H model with a plausible range of bivariate correlations, we executed simulation studies. We imitated the structure of the EDCAP data, an unbalanced cluster-randomized study, which included patient-level covariates (PSI1, PSI2, PSI3) and site-level covariates (Low, Mod, High). PSI1 and Low intensity intervention served as the reference levels. From each of the $m = 32$ sites, n_i patients are randomly generated from the Poisson distribution. Based on preliminary analysis of the EDCAP data indicated in Table 6, $\boldsymbol{\alpha}$ is chosen as (0.8, -1.5, -2.5, 1.0, 0.9), $\boldsymbol{\beta}$ is chosen as (1.3, 0.1, 0.4, -0.1, 0.1), $k = 2.6$, $\sigma_u = 0.6$, $\sigma_v = 0.2$, and ρ takes one of the following values (-0.1, -0.3, -0.5, -0.7). The number of replications is 1,000 for each of the four simulated settings.

In Table 7, we summarized one simulated dataset to show how well our simulated dataset captures the EDCAP data. We obtained the summary statistics similar to those for the EDCAP data of Table 1. In addition, Figure 10 confirms that the bed days of a simulated dataset reasonably represent those of the EDCAP data; the cumulative distribution of bed days by both are almost identical.

Results of the simulation studies, which are summarized in Tables 8 and 9, verify the performance of the proposed BLUP (REMQL) approach in the negative binomial H model. We reported the average bias, the relative bias to the true parameter (Percent), standard error (SE), mean square error (MSE), and coverage probability (CP) of the 95% confidence interval over 1000 replications for each value of ρ. For fixed effect estimates in the negative binomial part, ML and BLUP (REMQL) both give negligible biases. However, the BLUP (REMQL) fixed effects estimates in the logistic part yield somewhat larger biases than the ML estimates but still small bias (all percents $\leq$1.6). In addition, the estimate for the BLUP (REMQL) scale parameter has a larger bias than the ML estimate but still small bias (the largest percent=8.8). The BLUP (REMQL) estimates have much smaller biases than the ML estimates for the variance components $(\sigma_u, \sigma_v, \rho)$ of random effects (the smallest ratios: 8.2% vs. 3.0% for σ_u (Table 9(c)); 11.0% vs. 1.5% for σ_v (Table 8(a)); 66.3% vs. 5.3% for ρ (Table 8(b))). BLUP (REMQL) and ML both give similar SEs as well as MSEs except

the scale parameter. The proposed BLUP (REMQL) approach appears to have a slightly better or similar coverage probability except for the scale parameter across the scenarios considered.

In summary, simulation results demonstrate that the proposed estimation in the negative binomial H model with correlated random effects performs well for the linear predictors and variance components considered, but not for the scale parameter. We kept the same convergence criteria for both estimation methods; all replications converged for BLUP (REMQL), while some replications did not converge for ML (10/1000 replications at $\rho = -0.3$; 26/1000 replications at $\rho = -0.5$; 91/1000 replications at $\rho = -0.7$). BLUP (REMQL) runs in about 3/7 the time as ML. We used the SAS procedure NLMIXED to fit this model with ML (Appendix D), and used R to get the BLUP (REMQL) estimates (Appendix B).

4.6 DISCUSSION

In addition to a correlated random effects Poisson H model in Chapter 3, we have proposed a BLUP (REMQL) approach to obtain estimates in a correlated random effects negative binomial H model. We also applied this BLUP (REMQL) approach to account for potential problematic properties of bed days (excess zeros, clustering by site, and bivariate correlation between the binary and count components of the model) in the EDCAP study. The negative binomial H model accounted for the overdispersion from the heterogeneity between sites relative to a Poisson H model. Using a negative binomial H model with correlated random effects, we can simultaneously address an overall assessment of the intervention effect on two aspects of care, e.g., admission and LOS in the EDCAP study. The interventions in the EDCAP study were not initially designed to affect LOS in hospitals. Our results suggest that the intervention had a significant relationship only on the reduction of hospitalizations. Due to the correction of the scale parameter, we also estimated a lower negative bivariate correlation than in the Poisson H model with correlated random effects. This finding indicates that sites with low admission rates for low risk patients tend to have a somewhat shorter LOS for those admitted. In addition, the significant scale estimate indicates some

overdispersion relative to the Poisson. Using random effect predictions, we also identified sites showing a higher level of performance of the intervention.

As we found in a Poisson H model, this BLUP (REMQL) approach reduced the bias in variance components relative to the ML estimation in a negative binomial H model. Our simulation study showed that the BLUP (REMQL) approach provides estimates similar to ML for the linear predictors in the negative binomial part in our unbalanced study. The BLUP (REMQL) estimates of the scale parameter in the negative binomial part and the linear predictors in the logistic part had larger biases than the ML estimates. However, the proposed BLUP (REMQL) considerably reduced the running time relative to ML and had better convergence properties; ML did not converge for over 9% of the replications when $\rho = -0.7$.

Table 6: Correlated random effects negative binomial H model estimates based on (a) BLUP (REMQL) and (b) ML estimation

	(a) BLUP (REMQL)			(b) ML		
Parameter	**Estimate**	**SE**	**P-value**	**Estimate**	**SE**	**P-value**
			Logistic Part: Pr(Outpatient)			
Cons	0.79	0.25	<.01	0.80	0.24	<.01
PSI2	-1.47	0.13	<.001	-1.49	0.13	<.001
PSI3	-2.51	0.15	<.001	-2.54	0.15	<.001
Mod	0.97	0.30	<.01	0.98	0.29	<.01
High	0.91	0.30	<.01	0.93	0.29	<.01
$\boldsymbol{\sigma}_u$	0.58			0.54		
			Negative Binomial Part: Inpatient Bed Days			
Cons	1.30	0.10	<.001	1.29	0.10	<.001
PSI2	0.12	0.09	.18	0.12	0.09	.18
PSI3	0.38	0.09	<.001	0.39	0.09	<.001
Mod	-0.04	0.10	.68	-0.04	0.10	.68
High	0.03	0.11	.77	0.03	0.10	.75
k	2.71	0.26	<.001	2.56	0.26	<.001
$\boldsymbol{\sigma}_v$	0.17			0.15		
$\boldsymbol{\rho}$	**-0.10**			**-0.12**		
-2* ℓ		5959.5			6066.0	
AIC		5987.5			6094.0	

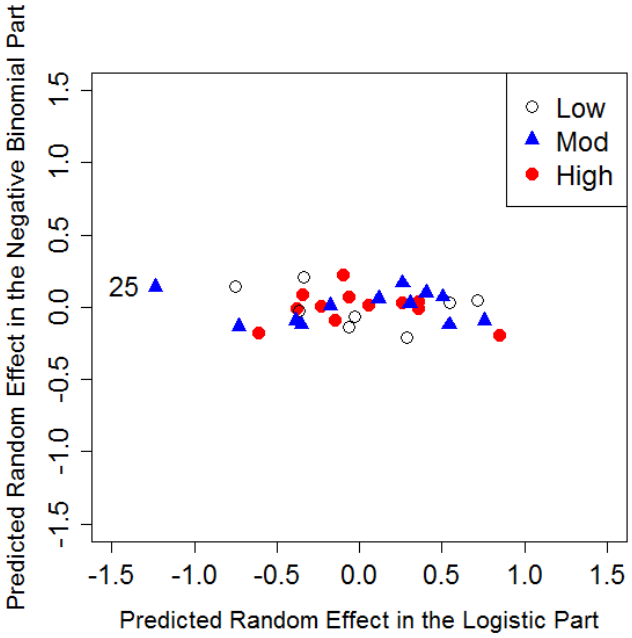

Figure 8: Site specific predicted random effects for the logistic and negative binomial parts of the negative binomial H model for the low (○), moderate (▲), and high (●) intensity intervention sites

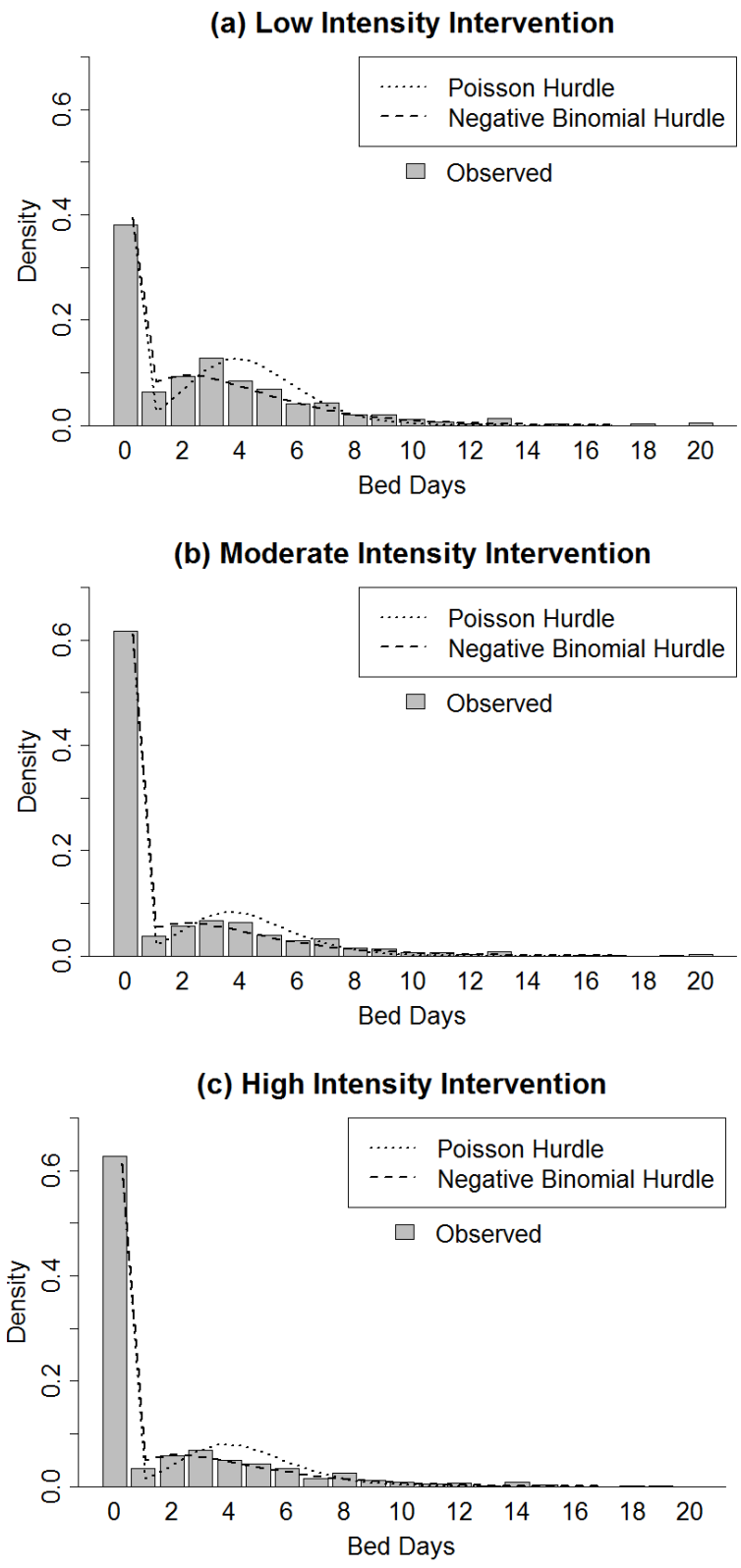

Figure 9: Observed vs predicted distribution of bed days by intervention arm

Table 7: Probability of outpatient, mean and median of inpatients bed days, and mean and median of overall bed days by PSI risk class and by intervention arm for one simulated dataset (N=1,823)

			Bed Days	
	n	**Pr(Outpatient)**	**Inpatients Mean(Median)**	**Overall Mean(Median)**
PSI risk class				
1	654	0.79	3.9 (4.0)	0.8 (0.0)
2	644	0.53	4.8 (4.0)	2.3 (0.0)
3	525	0.33	6.0 (5.0)	4.0 (3.0)
Intervention				
Low	532	0.42	5.0 (4.0)	2.9 (2.0)
Mod	650	0.59	4.6 (4.0)	1.9 (0.0)
High	641	0.66	6.0 (5.0)	2.1 (0.0)
Overall		0.57	5.2 (4.0)	2.2 (0.0)

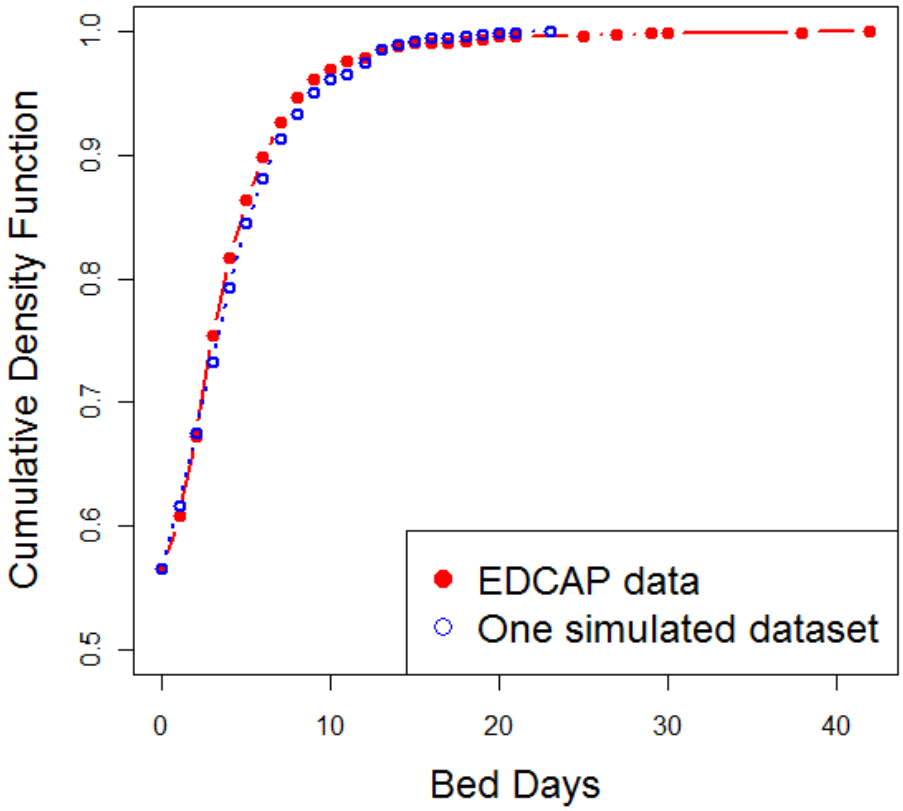

Figure 10: Cumulative density function of bed days by EDCAP data (•) and one simulated dataset (◦) with $\rho = -0.1$

Table 8: Simulation results using correlated random effects negative binomial H model based on ML and BLUP (REMQL) estimation with 1000 replications and a plausible range of bivariate correlation ($\rho = -0.1, -0.3$)

		Bias(Percent*)		SE		MSE		CP	
Parameter	**True**	**ML**	**REMQL**	**ML**	**REMQL**	**ML**	**REMQL**	**ML**	**REMQL**
(a) $\rho = -0.1$				**Logistic Part: Pr(Outpatient)**					
α_0 : **Cons**	0.8	-0.007(0.9)	-0.016(2.0)	0.247	0.256	0.136	0.139	0.93	0.94
α_1 : **PSI2**	-1.5	-0.007(0.5)	0.010(0.7)	0.130	0.129	0.035	0.034	0.94	0.94
α_2 : **PSI3**	-2.5	-0.008(0.3)	0.021(0.8)	0.150	0.148	0.046	0.045	0.96	0.95
α_3 : **Mod**	1.0	0.008(0.8)	-0.004(0.4)	0.305	0.317	0.204	0.209	0.93	0.94
α_4 : **High**	0.9	0.025(2.8)	0.014(1.6)	0.306	0.318	0.200	0.204	0.93	0.94
σ_u	0.6	-0.039(6.5)	-0.009(1.5)						
				Negative Binomial Part: Inpatient Bed Days					
β_0 : **Cons**	1.3	-0.002(0.2)	0.003(0.2)	0.107	0.110	0.025	0.025	0.94	0.94
β_1 : **PSI2**	0.1	0.002(2.0)	0.001(1.0)	0.085	0.083	0.015	0.014	0.95	0.95
β_2 : **PSI3**	0.4	0.004(1.0)	0.000(-)	0.084	0.082	0.014	0.014	0.95	0.95
β_3 : **Mod**	-0.1	-0.001(1.0)	-0.001(1.0)	0.114	0.119	0.029	0.030	0.92	0.93
β_4 : **High**	0.1	0.001(1.0)	0.000(-)	0.113	0.118	0.028	0.029	0.93	0.94
k	2.6	0.038(1.5)	0.229(8.8)	0.284	0.286	0.169	0.238	0.96	0.90
σ_v	0.2	-0.022(11.0)	-0.003(1.5)						
ρ	-0.1	0.063(63.0)	-0.003(3.0)						
(b) $\rho = -0.3$				**Logistic Part: Pr(Outpatient)**					
α_0 : **Cons**	0.8	-0.001(0.1)	-0.010(1.3)	0.244	0.253	0.130	0.133	0.95	0.96
α_1 : **PSI2**	-1.5	-0.011(0.7)	0.006(0.4)	0.130	0.129	0.034	0.033	0.95	0.95
α_2 : **PSI3**	-2.5	-0.002(0.1)	0.026(1.0)	0.150	0.147	0.045	0.044	0.96	0.95
α_3 : **Mod**	1.0	0.003(0.3)	-0.008(0.8)	0.302	0.314	0.197	0.202	0.93	0.94
α_4 : **High**	0.9	0.008(0.9)	-0.002(0.2)	0.301	0.313	0.192	0.197	0.93	0.94
σ_u	0.6	-0.045(7.5)	-0.016(2.7)						
				Negative Binomial Part: Inpatient Bed Days					
β_0 : **Cons**	1.3	-0.002(0.2)	0.004(0.3)	0.107	0.109	0.025	0.025	0.93	0.93
β_1 : **PSI2**	0.1	0.002(2.0)	0.001(1.0)	0.085	0.083	0.015	0.014	0.95	0.95
β_2 : **PSI3**	0.4	-0.002(0.5)	-0.005(1.3)	0.084	0.082	0.015	0.014	0.94	0.93
β_3 : **Mod**	-0.1	0.001(1.0)	0.001(1.0)	0.114	0.120	0.027	0.028	0.93	0.95
β_4 : **High**	0.1	-0.001(1.0)	-0.002(2.0)	0.112	0.118	0.027	0.028	0.93	0.95
k	2.6	0.036(1.4)	0.219(8.4)	0.282	0.284	0.159	0.222	0.96	0.91
σ_v	0.2	-0.023(11.5)	-0.003(1.5)						
ρ	-0.3	0.199(66.3)	0.016(5.3)						

* indicates the relative bias to the true parameter, which is calculated by 100*abs (Bias/True).

For ML, 10 replications did not converge when $\rho = -0.3$.

Table 9: Simulation results using correlated random effects negative binomial H model based on ML and BLUP (REMQL) estimation with 1000 replications and a plausible range of bivariate correlation ($\rho = -0.5, -0.7$)

		Bias(Percent*)		**SE**		**MSE**		**CP**	
Parameter	**True**	**ML**	**REMQL**	**ML**	**REMQL**	**ML**	**REMQL**	**ML**	**REMQL**
(c) $\boldsymbol{\rho} = -0.5$				**Logistic Part: Pr(Outpatient)**					
$\boldsymbol{\alpha}_0$: **Cons**	0.8	-0.016(2.0)	-0.024(3.0)	0.242	0.253	0.127	0.132	0.94	0.95
$\boldsymbol{\alpha}_1$: **PSI2**	-1.5	0.002(0.1)	0.018(1.2)	0.129	0.128	0.035	0.035	0.93	0.92
$\boldsymbol{\alpha}_2$: **PSI3**	-2.5	-0.002(0.1)	0.024(1.0)	0.148	0.146	0.044	0.044	0.96	0.95
$\boldsymbol{\alpha}_3$: **Mod**	1.0	0.003(0.3)	-0.007(0.7)	0.300	0.314	0.192	0.199	0.93	0.94
$\boldsymbol{\alpha}_4$: **High**	0.9	0.010(1.1)	0.001(0.1)	0.298	0.312	0.193	0.200	0.93	0.95
$\boldsymbol{\sigma}_u$	0.6	-0.049(8.2)	-0.018(3.0)						
				Negative Binomial Part: Inpatient Bed Days					
$\boldsymbol{\beta}_0$: **Cons**	1.3	0.004(0.3)	0.008(0.6)	0.107	0.11	0.024	0.025	0.94	0.94
$\boldsymbol{\beta}_1$: **PSI2**	0.1	0.003(3.0)	0.003(3.0)	0.083	0.082	0.014	0.014	0.95	0.95
$\boldsymbol{\beta}_2$: **PSI3**	0.4	0.001(0.3)	-0.002(0.5)	0.083	0.081	0.014	0.013	0.96	0.95
$\boldsymbol{\beta}_3$: **Mod**	-0.1	-0.006(6.0)	-0.005(5.0)	0.114	0.121	0.028	0.029	0.94	0.95
$\boldsymbol{\beta}_4$: **High**	0.1	-0.007(7.0)	-0.008(8.0)	0.112	0.119	0.026	0.028	0.94	0.96
k	2.6	0.051(2.0)	0.221(8.5)	0.280	0.280	0.169	0.222	0.94	0.90
$\boldsymbol{\sigma}_v$	0.2	-0.020(10.0)	0.000(-)						
$\boldsymbol{\rho}$	-0.5	0.326(65.2)	0.025(5.0)						
(d) $\boldsymbol{\rho} = -0.7$				**Logistic Part: Pr(Outpatient)**					
$\boldsymbol{\alpha}_0$: **Cons**	0.8	0.003(0.4)	-0.004(0.5)	0.243	0.252	0.130	0.134	0.93	0.95
$\boldsymbol{\alpha}_1$: **PSI2**	-1.5	0.003(0.2)	0.016(1.1)	0.130	0.129	0.034	0.034	0.95	0.94
$\boldsymbol{\alpha}_2$: **PSI3**	-2.5	0.003(0.1)	0.027(1.1)	0.149	0.147	0.045	0.044	0.95	0.94
$\boldsymbol{\alpha}_3$: **Mod**	1.0	0.002(0.2)	-0.007(0.7)	0.302	0.314	0.200	0.206	0.93	0.94
$\boldsymbol{\alpha}_4$: **High**	0.9	0.002(0.2)	-0.006(0.7)	0.300	0.312	0.195	0.201	0.94	0.95
$\boldsymbol{\sigma}_u$	0.6	-0.047(7.8)	-0.016(2.7)						
				Negative Binomial Part: Inpatient Bed Days					
$\boldsymbol{\beta}_0$: **Cons**	1.3	-0.001(0.1)	0.002(0.2)	0.107	0.110	0.023	0.024	0.94	0.95
$\boldsymbol{\beta}_1$: **PSI2**	0.1	-0.003(3.0)	-0.003(3.0)	0.084	0.082	0.014	0.014	0.94	0.94
$\boldsymbol{\beta}_2$: **PSI3**	0.4	-0.001(0.3)	-0.004(1.0)	0.083	0.081	0.014	0.013	0.94	0.94
$\boldsymbol{\beta}_3$: **Mod**	-0.1	0.000(-)	0.000(-)	0.115	0.122	0.028	0.029	0.94	0.95
$\boldsymbol{\beta}_4$: **High**	0.1	0.004(4.0)	0.003(3.0)	0.113	0.119	0.027	0.028	0.93	0.94
k	2.6	0.068(2.6)	0.204(7.8)	0.284	0.280	0.179	0.212	0.95	0.91
$\boldsymbol{\sigma}_v$	0.2	-0.018(9.0)	0.002(1.0)						
$\boldsymbol{\rho}$	-0.7	0.455(65.0)	0.035(5.0)						

* indicates the relative bias to the true parameter, which is calculated by 100*abs (Bias/True).

For ML, 26 replications did not converge when $\rho = -0.5$ and 91 replications did not converge when $\rho = -0.7$.

5.0 DISCUSSION

In health service studies, bed days could be a relevant metric to quantify efficiency of care. This dissertation, which focused on an efficient statistical modeling for bed days, propose a BLUP (REMQL) approach to estimate parameters of Poisson/negative binomial H models with correlated random effects. These models appropriately account for excess zeros, possible clustering by site (in multi-site studies), and possible bivariate correlation between the binary and count components of these models. In addition, the correlated random effects negative binomial H model allows the overdispersion relative to the Poisson H model. Advantages of these models include an overall assessment of the effect on an intervention for two aspects of care, e.g., admission and LOS in the EDCAP study. The interventions in the EDCAP study were designed to influence the admission decision and increase performance of recommended processes of care in the Emergency Department (ED); our results showed that the intensity of the intervention was significantly associated with reduced hospitalization but not with LOS. In addition, the low negative bivariate correlation indicates some small tendency for sites with relatively low admission rates for low risk patients to have a somewhat shorter LOS for those admitted.

Correlated random effects allow the characterization of the relative amount of site variation with respect to both the hospitalization decision and LOS. For the scenarios considered, our simulation study indicated that the BLUP (REMQL) approach provides less biased estimates of variance components than the ML and gives estimates similar to the ML for the linear predictors in the count part. However, the BLUP (REMQL) also yields somewhat more biased estimates of the linear predictors in the logistic part and a larger bias in the estimated scale parameter relative to the ML. We need to account for these issues to improve our BLUP (REMQL) approach in correlated random effects H models. However,

BLUP (REMQL) estimation still is attractive because it ran faster than the ML and provided better convergence properties.

In summary, the proposed BLUP (REMQL) estimation in these H models appears to be promising. Although this dissertation research focuses on a mix of excess zeros and one level of clustering (by site), in public health applications, excess zeros can be observed at multiple levels. For example, the data in the motivating example were clustered by provider as well as site. The computational advantages of the BLUP (REMQL) approach may become more apparent in more complex models, such as a 3-level model. Our current research deals with model estimation, but it also acknowledges robustness and influence assessment when the relative performance of sites is of interest. Another future research direction is exploring influence diagnostics for a correlated random effects hurdle model at both the site and patient levels. These diagnostics could be derived from case-deletion or local influence in the GLMM, both of which have been developed but not implemented for a correlated random effects H model.

APPENDIX A

BLUP (REMQL) ESTIMATION R CODE IN THE CORRELATED RANDOM EFFECTS POISSON HURDLE MODEL

Steps for programming in R

1. **Parameter setting**
2. **Get initial values for the parameters**

   $\boldsymbol{\alpha}_0 \leftarrow$ **Random effects logistic regression**

   $\boldsymbol{\beta}_0 \leftarrow$ **Random effects Poisson regression**

   $\boldsymbol{\phi}_0 = (\sigma_{u0}, \sigma_{v0}, \rho_0) \leftarrow$ **Random effects logistic regression and random effects Poisson regression**
3. **Estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using N-R algorithm**
4. **Get $\boldsymbol{\phi} = (\sigma_u, \sigma_v, \rho)$ using N-R algorithm to solve the REMQL estimating equations**

We will only present the R code for the EDCAP study as a simple example.

The following six sub-functions should be defined before running the "REPoissonH" function:

---

1) wreml.logit: function for the logistic regression in the GLMM setting

```
wreml.logit = function(y,x,z,alfa1,yu1,sig1,famaly="logistic",epsilon=1e-3)
{
M = ncol(z);n = length(y); sigu.2 = sig1^2
X = cbind(1,x)
p1 = ncol(X)
zero1 = matrix(0,ncol=p1,nrow=M)
X1 = rbind(X,zero1)
Z = rbind(z,diag(M))
XX = cbind(X1,Z)
itmax = 1000
Alfa0 = c(alfa1,yu1)
alfa0 = alfa1
yu0 = yu1;flag = 0
for(iter in 1:itmax)
{
for(it in 1:itmax)
{
theta = as.vector(X%*%alfa0+z%*%yu0)
w1 = exp(theta)/(1+exp(theta))^2
w = c(w1,rep(1/sigu.2,M))
mu = exp(theta)/(1+exp(theta))
ply = c(as.vector(t(X)%*%(y-mu)),as.vector(t(z)%*%(y-mu)-yu0/sigu.2))
w = t(matrix(rep(w,(p1+M)),ncol=(p1+M)))
V1 = (t(XX)*w)%*%XX
V = solve(V1)
Alfa = Alfa0 + V %*% ply
alfa = Alfa[1:p1]
yu = Alfa[(p1+1):(p1+M)]
if(max(abs(Alfa-Alfa0))<epsilon)  flag = 1;break
Alfa0 = Alfa; alfa0 = alfa; yu0 = yu
```

```
}
if(!flag) break
nsigu.2 = as.vector(t(yu)%*%yu + sum(diag(V)[(p1+1):(p1+M)]))/M
if(abs(nsigu.2-sigu.2)<epsilon) flag2 = 1;break
sigu.2 = nsigu.2
}
if(flag2) result = list(alfa=alfa,yu=yu, sig1=sqrt(nsigu.2), IV=V, prob=mu)
else stop("error:not reach the convergence")
}
```

2) wreml.poi: function for GLMM of truncated Poisson regression

```
wreml.poi = function(y, zk , x, z, beta1, va1, sig2, fam="Poisson", epsilon=1e-3)
{
M = ncol(z);n = length(y); sigv.2 = sig2^2
X = cbind(1,x);p1 = ncol(X)
zero1 = matrix(0,ncol=p1,nrow=M)
X1 = rbind(X,zero1)
Z = rbind(z,diag(M))
XX = cbind(X1,Z)
itmax = 1000;
Alfa0 = c(beta1,va1)
beta0 = beta1 ; va0 = va1
flag = 0
for(iter in 1:itmax)
{
for(it in 1:itmax)
{
theta = as.vector(X%*%beta0+z%*%va0)
lamda = exp(theta)
w1 = (1-zk)*(lamda*((1-exp(-lamda)-lamda*exp(-lamda))/(1-exp(-lamda))^2))
```

```
w = c(w1,rep(1/sigv.2,M))
ply = c(as.vector(t(X)%*%((1-zk)*(y+(-lamda/(1-exp(-lamda))))))),
as.vector(t(z)%*%((1-zk)*(y+(-lamda/(1-exp(-lamda)))))-va0/sigv.2))
w = t(matrix(rep(w,(p1+M)),ncol=(p1+M)))
V1 = (t(XX)*w)%*%XX
V = solve(V1)
Alfa = Alfa0 + V%*%ply
beta = Alfa[1:p1]
va = Alfa[(p1+1):(p1+M)]
if(max(abs(Alfa-Alfa0))<epsilon)  flag = 1;break
Alfa0 = Alfa; beta0 = beta; va0 = va
}
if(!flag) break
nsigv.2 = as.vector(t(va)%*%va + sum(diag(V)[(p1+1):(p1+M)]))/M
if(abs(nsigv.2 - sigv.2)<epsilon)flag2 = 1;break
sigv.2 = nsigv.2
}
if(flag2) result = list(beta = beta, va = va, sig2=sqrt(nsigv.2), IV=V, lamda = lamda)
else stop("error: not reach the convergence")
}
```

3) getIA: function to get IA using IAi

```
getIA = function(sigu,sigv,rho,m)
{
IAi = solve(matrix(c(sigu^2,sigu*sigv*rho,sigu*sigv*rho,sigv^2),ncol=2))
IA = diag(2*m)
for (i in 1:m)
IA[2*(i-1)+1:2,2*(i-1)+1:2] = IAi
IA
}
```

4) L1L2L3: function to get L1, L2, L3

```
L1L2L3 = function(x,m)
{
L3 = diag(x)%*%rep(c(0,1),m)
L1 = diag(x)%*%rep(c(1,0),m)
L2 = counter = 0
for(j in 1:m)
{
L2 = L2+x[counter+2,counter+1]
counter = counter+2
}
list(L1=L1,L2=L2,L3=L3)
}
```

5) sigu.sigv.rho : function to get the estimates for sigu, sigv, rho using the N-R method

```
sigu.sigv.rho = function(sigu,sigv,rho,L1,L2,L3,m)
{
vect = c(sigu,sigv,rho)
for(h in 1:10)
{ fa = m*(1-rhoˆ2)*(sigu*sigv)ˆ2-sigvˆ2*L1+rho*sigu*sigv*L2
fb = m*(1-rhoˆ2)*(sigu*sigv)ˆ2-siguˆ2*L3+rho*sigu*sigv*L2
fc = -m*(1-rhoˆ2)*(sigu*sigv)ˆ2*rho+rho*sigvˆ2*L1+rho*siguˆ2*L3-sigu*sigv*(rhoˆ2+1)*L2
fa1 = 2*m*(1-rhoˆ2)*sigu*sigvˆ2+rho*sigv*L2
fa2 = 2*m*(1-rhoˆ2)*siguˆ2*sigv+rho*sigu*L2-2*sigv*L1
fa3 = -2*m*rho*(sigu*sigv)ˆ2+sigu*sigv*L2
fb1 = 2*m*(1-rhoˆ2)*sigvˆ2*sigu+rho*sigv*L2-2*sigu*L3
fb2 = 2*m*(1-rhoˆ2)*siguˆ2*sigv+rho*sigu*L2
fb3 = -2*m*rho*(sigu*sigv)ˆ2+sigu*sigv*L2
fc1 = -2*m*(1-rhoˆ2)*sigu*sigvˆ2*rho+2*rho*sigu*L3-sigv*(rhoˆ2+1)*L2
```

```
fc2 = -2*m*(1-rhoˆ2)*sigv*siguˆ2*rho+2*rho*sigv*L1-sigu*(rhoˆ2+1)*L2
fc3 = -m*(1-3*rhoˆ2)*(sigu*sigv)ˆ2+sigv2̂*L1+siguˆ2*L3-sigu*sigv*2*rho*L2
F2 = solve(t(cbind(c(fa1,fa2,fa3),c(fb1,fb2,fb3),c(fc1,fc2,fc3))))
F1 = c(fa,fb,fc)
nvect = vect-F2%*%F1
if(max(abs(nvect-vect))<0.001) break
vect = nvect
sigu = vect[1]; sigv = vect[2]; rho = vect[3]
}
vect
}
```

6) biv.hessen: function to get the first derivatives and V

```
biv.hessen = function(y,r,H,alfa,yu,beta,va,IA,W,X,R)
{
# Data set-up
p1 = ncol(W)
p2 = ncol(X)
n = nrow(W)
zeroy = ifelse(y==0,1,0)
ksi = as.vector(W %*% alfa + R %*% yu)
eta = as.vector(X %*% beta + R %*% va)
# The first derivatives
dl.ksi = zeroy - exp(ksi)/(1+exp(ksi))
dl.eta = (1-zeroy)*(y - exp(eta)/(1-exp(-exp(eta))))
dl.alfa = t(W) %*% dl.ksi
dl.beta = t(X) %*% dl.eta
dl.r = t(H) %*% c(t(R)%*%dl.ksi, t(R)%*%dl.eta)-as.vector(IA%*%r)
f = c(dl.alfa, dl.beta, dl.r)
# −The second derivatives
```

```
    d2l.ksi2 = exp(ksi)/(1+exp(ksi))^ 2
    d2l.eta2 = (1-zeroy)*((exp(eta)*(1-exp(-exp(eta))-exp(eta)*exp(-exp(eta))))/(1-
    exp(-exp(eta)))^ 2)
    d2l.ksi.eta = rep(0,n)
    I11 = t(W*d2l.ksi2)%*%W
    I12 = t(W*d2l.ksi.eta)%*%X
    I13 = cbind(t(W*d2l.ksi2)%*%R, t(W*d2l.ksi.eta)%*%R)%*%H
    I22 = t(X*d2l.eta2)%*%X
    I23 = cbind(t(X*d2l.ksi.eta)%*%R, t(X*d2l.eta2)%*%R)%*%H
    I33 = t(H) %*% rbind(cbind(t(R*d2l.ksi2)%*%R, t(R*d2l.ksi.eta)%*%R),
    cbind(t(R*d2l.ksi.eta)%*%R, t(R*d2l.eta2)%*%R)) %*% H + IA
    V = rbind(cbind(I11,I12,I13),cbind(t(I12),I22,I23),cbind(t(I13),t(I23),I33))
    list(f=f, V=V)
    }
```

The main function of BLUP (REMQL) estimation for the correlated random effects Poisson H model can be defined as follows:

```
REPoissonH = function(data)
{
# Data set-up
    y = data[,1]
    zeroy = ifelse(y==0,1,0)
    x.l = as.matrix(data[,3:dim(data)[2]])
    n = length(y); m = max(data[,2])
    R = matrix(0, ncol = m, nrow = n)
        for (i in 1:m)
        R[,i] = ifelse(data[,2] == i, 1, 0)
    W = cbind(1,x.l)
```

```
    X = cbind(1,x.l)
    p1 = ncol(W)
    p2 = ncol(X)
# Calculate H
    a = rbind(diag(m),diag(0,m))
    b = rbind(diag(0,m),diag(m))
    H = cbind(a[,1],b[,1])
    for (i in 2:m)
    H = cbind(H,a[,i],b[,i])
# Initial values for the parameters
    ct0 = list(epsilon = 0.001, maxit = 50, trace = F)
    alfa1 = coef(glm(zeroy~x.l, family = binomial(link = "logit"),
        na.action = na.omit, control = ct0))
    beta1 = coef(glm(y~x.l, family = poisson(link = "log"),
        na.action = na.omit, control = ct0))
    yu = rep(0,m)
    va = rep(0,m)
    sigu1 = 0.5
    sigv1 = 0.5
    rho = 0
    glm.logit = wreml.logit(zeroy, x.l, R, alfa1, yu, sigu1)
    alfa = as.vector(glm.logit$alfa)
    sigu = glm.logit$sig1
    u = glm.logit$yu
    glm.poi = wreml.poi(y, zeroy, x.l, R, beta1, va, sigv1)
    beta = as.vector(glm.poi$beta)
    sigv = glm.poi$sig2
    v = glm.poi$va
    r = t(H) %*% c(u,v)
    coef = c(alfa,beta,r)
```

```
# Loop control
    flag = 0; epsilon = 0.001; itmax = 1000
# Begin of it loop
    for (it in 1:itmax)
    {
    IA = getIA(sigu,sigv,rho,m)
# Begin of iter loop
        for (iter in 1:1000)
        {
        f.V = biv.hessen(y,r,H,alfa,u,beta,v,IA,W,X,R)
        f = f.V$f
        V = f.V$V
        IV = try(solve(V))
        coef0 = coef + IV %*% f
        test = try(max(abs((coef-coef0))) < epsilon)
        if(test) flag = 1;break
        coef = coef0
        alfa = coef[1:p1]; beta = coef[(p1+1):(p1+p2)]; r = coef[(p1+p2+1):(p1+p2+2*m)]
        r0 = H %*% r; u = r0[1:m]; v = r0[(m+1):(2*m)]
        }
# End of iter loop
    if(!flag) break; flag = 0
# Get L1,L2,L3
    S = IV[(p1+p2+1):(p1+p2+2*m), (p1+p2+1):(p1+p2+2*m)]
    x = S + r %*% t(r)
    L = L1L2L3(x,m)
    L1 = L$L1; L2 = L$L2; L3 = L$L3
# Get sigu, sigv, rho
    vect = sigu.sigv.rho(sigu,sigv,rho,L1,L2,L3,m)
    sigu0 = vect[1];sigv0 = vect[2];rho0 = vect[3]
```

```
    if(max(abs(c((sigu-sigu0),(sigv-sigv0),(rho-rho0)))) < epsilon)
    {flag = 1; break}
    sigu = sigu0; sigv = sigv0; rho = rho0;
    }
# End of it loop
    if(!flag) result = list(NULL)
    else
# Get standard errors for the parameters
    {
    se = sqrt(diag(IV))
    se.alfa = se[1:p1]
    se.beta = se[(p1+1):(p1+p2)]
# Calculate the loglikelihood
    ksi = W%*%alfa + R%*%u; eta = X%*%beta + R%*%v;
    p = exp(ksi)/(1+exp(ksi)); mu = exp(eta)
    loglik = sum(zeroy*log(p) + (1-zeroy)*log(1-p) + (1-zeroy)*(-mu + y*log(mu)
    - log(gamma(y+1)) - log(1-exp(-mu))))
    result = list(alfa=alfa,beta=beta,sigu=sigu,sigv=sigv,rho=rho,se.alfa=se.alfa,
    se.beta=se.beta,loglik=loglik,u=u,v=v)
    }
# Output
    result
}
```

Here is the sample of EDCAP data:

stnum	raceenr	physcid	physcgid	site	ct	trt	los	diedhosp	psiclass	psihigh	inpat
11000213	1	11001	711001	11	0	1	2	0	1.0	0	1
11000238	2	11001	711001	11	0	1	7	0	3.5	1	1
11000417	1	11001	711001	11	0	1	0	0	2.0	0	0
11000427	1	11001	711001	11	0	1	7	0	5.0	1	1

11000431 1 11001 711001 11 0 1 0 0 1.0 0 0

```
# Import the EDCAP data in R
    data = read.table("C://edcap_3112.txt", header=T, as.is=T)
# Zero LOS for inpatients was counted as 1 day
    for (i in 1:(length(data[,8])))
    {
    if(data[i,8] == 0.5) data[i,8] = 1
    }
# Keep only low risk PSI patients (N=1,877)
    data = subset(data, psihigh == 0)
    attach(data)
    names(data)
    psi.2 = as.numeric(1<psiclass & psiclass<3)
    psi.3 = as.numeric(psiclass>2)
    trt.2 = as.numeric(1<trt & trt<3)
    trt.3 = as.numeric(trt>2)
    edcap = cbind(los, site, trt.2, trt.3, psi.2, psi.3)
    Model = REPoissonH(edcap)
```

APPENDIX B

BLUP (REMQL) ESTIMATION R CODE IN THE CORRELATED RANDOM EFFECTS NEGATIVE BINOMIAL HURDLE MODEL

Steps for programming in R

1. **Parameter setting**
2. **Get initial values for the parameters**
   $\boldsymbol{\alpha}_0 \leftarrow$ **Random effects logistic regression**
   $\boldsymbol{\beta}_0$ **and** $k_0 \leftarrow$ **Random effects negative binomial regression**
   $\boldsymbol{\phi}_0 = (\sigma_{u0}, \sigma_{v0}, \rho_0) \leftarrow$ **Random effects logistic regression and random effects negative binomial regression**
3. **Estimate** $\boldsymbol{\alpha}$ **and** $\boldsymbol{\beta}$ **using N-R algorithm**
4. **Get** $\boldsymbol{\phi} = (\sigma_u, \sigma_v, \rho)$ **using N-R algorithm to solve the REMQL estimating equations**
5. **Get** k **using profile loglikelihood**

We will only present the R code for the EDCAP study as a simple example.

The following three sub-functions should be defined before running the "REnbH" function (repeated sub-functions, which are presented at the previous R code for the Poisson hurdle model, were omitted):

---

1) wreml.tnb: function for the truncated negative binomial regression in the GLMM setting

```
wreml.tnb = function(y, zk , x, z, beta1, va1, sig2, k, p, fam="NB", epsilon=1e-3)
{
M = ncol(z);n = length(y); sigv.2 = sig2^2
X = cbind(1,x);p1 = ncol(X)
zero1 = matrix(0,ncol=p1,nrow=M)
X1 = rbind(X,zero1)
Z = rbind(z,diag(M))
XX = cbind(X1,Z)
itmax = 1000;
Alfa0 = c(beta1,va1)
beta0 = beta1 ; va0 = va1;
flag = 0
for(iter in 1:itmax)
{
for(it in 1:itmax)
{
theta = as.vector(X%*%beta0+z%*%va0)
lamda = exp(theta)
t = k /(k+lamda)
w1 = (1-zk)*t*(1-t)*(y-(k^2*(1-t)*t^(k-1)-k*(1-t^k))/(1-t^k)^2)
w = c(w1,rep(1/sigv.2,M))
ply = c(as.vector(t(X)%*%((1-zk)*(y*t-(k*(1-t))/(1-t^k)))),
as.vector(t(z)%*%((1-zk)*(y*t-(k*(1-t))/(1-t^k)))-va0/sigv.2))
w = t(matrix(rep(w,(p1+M)),ncol=(p1+M)))
V1 = (t(XX)*w)%*%XX
V = solve(V1)
Alfa = Alfa0 + V%*%ply
beta = Alfa[1:p1]
```

```
va = Alfa[(p1+1):(p1+M)]
if(max(abs(Alfa-Alfa0))<epsilon) { flag = 1;break}
Alfa0 = Alfa; beta0 = beta; va0 = va;
}
if (!flag) break
nsigv.2 = as.vector(t(va)%*%va + sum(diag(V)[(p1+1):(p1+M)]))/M
k0 = agetk.ml(y, lamda, zk, p)
if (max(abs(c(nsigv.2,k0)-c(sigv.2,k)))<epsilon) { flag2 = 1; break}
sigv.2 = nsigv.2; k = k0
}
if(flag2)
{
# Standard error for the k
c = max(y)
t = k0/(k0+lamda)
f0 = table(y[y>0])
f = rep(0, c)
f[as.numeric(names(f0))] = f0
tot = sum(f0)
f = tot + f - cumsum(f)
i = sum(f/(k0+1:c-1))
ii = - sum(f/(k0+1:c-1)^2)
B = log(t)+(1-t)
B1 = (1-t)^2/k0
ep1 = t^k0
ep2 = (1-ep1)^2
w33 = (-sum((1-zk)*(B1*(1-ep1)+B^2*ep1)/ep2+(1-zk)*y*(t/k0)^2))-ii
se.k = sqrt(1/w33)
result = list(beta = beta, va = va, k = k0, sig2=sqrt(nsigv.2), se.k=se.k, IV=V)
}
```

```
else stop("error: not reach the convergence")
}
```

2) agetk.ml: function to get the estimate for the k

```
agetk.ml = function(y, mu, ZK, p)
{
loglik = function(th,y,mu,p,ZK)
{
k = exp(th)
t = k/(k+mu)
(sum( ZK*log(p)+(1-ZK)*log(1-p)+(1-ZK)*(log(gamma(y+k)/gamma(y+1)/gamma(k))
+k*log(t)+y*log(1-t)-log(1-tˆk)) ))
}
objm = optimize(loglik, lower = -8, upper = 5, y=y, mu=mu, ZK=ZK, p=p, maxi-
mum=T)
th = objm$maximum
exp(th)
}
```

3) biv.hessen: function to get the first derivatives and V

```
biv.hessen = function(y,r,H,alfa,yu,beta,va,IA,W,X,R,k)
{
# Data set-up
p1 = ncol(W)
p2 = ncol(X)
n = nrow(W)
zeroy = ifelse(y==0,1,0)
ksi = as.vector(W %*% alfa + R %*% yu)
eta = as.vector(X %*% beta + R %*% va)
mu = exp(eta)
```

```
t = k/(k+mu)
# The first derivatives
dl.ksi = zeroy - exp(ksi)/(1+exp(ksi))
dl.eta = (1-zeroy)*(y*t-(k*(1-t))/(1-t^k))
dl.alfa = t(W) %*% dl.ksi
dl.beta = t(X) %*% dl.eta
dl.r = t(H) %*% c(t(R)%*%dl.ksi, t(R)%*%dl.eta)-as.vector(IA%*%r)
f = c(dl.alfa, dl.beta, dl.r)
# −The second derivatives
d2l.ksi2 = diag( exp(ksi)/(1+exp(ksi))^2 )
d2l.eta2 = diag( (1-zeroy)*t*(1-t)*(y-(k^2*(1-t)*t^(k-1)-k*(1-t^k))/(1-t^k)^2) )
d2l.ksi.eta = diag( rep(0,n) )
I11 = t(W)%*%d2l.ksi2%*%W
I12 = t(W)%*%d2l.ksi.eta%*%X
I13 = cbind(t(W)%*%d2l.ksi2%*%R, t(W)%*%d2l.ksi.eta%*%R)%*%H
I22 = t(X)%*%d2l.eta2%*%X
I23 = cbind(t(X)%*%d2l.ksi.eta%*%R, t(X)%*%d2l.eta2%*%R)%*%H
I33 = t(H) %*% rbind(cbind(t(R)%*%d2l.ksi2%*%R, t(R)%*%d2l.ksi.eta%*%R),
cbind(t(R)%*%d2l.ksi.eta%*%R, t(R)%*%d2l.eta2%*%R)) %*% H + IA
V = rbind(cbind(I11,I12,I13),cbind(t(I12),I22,I23),cbind(t(I13),t(I23),I33))
list(f=f, V=V)
}
```

The main function of BLUP (REMQL) estimation for the correlated random effects negative binomial H model is presented as follows:

```
REnbH = function(data)
{
# Data set-up
```

```
    y = data[,1]
    zeroy = ifelse(y==0,1,0)
    x.l = as.matrix(data[,3:dim(data)[2]])
    n = length(y); m = max(data[,2])
    R = matrix(0, ncol = m, nrow = n)
        for (i in 1:m)
        R[,i] = ifelse(data[,2] == i, 1, 0)
    W = cbind(1,x.l)
    X = cbind(1,x.l)
    p1 = ncol(W)
    p2 = ncol(X)
# Calculate H
    a = rbind(diag(m),diag(0,m))
    b = rbind(diag(0,m),diag(m))
    H = cbind(a[,1],b[,1])
        for (i in 2:m)
        H = cbind(H,a[,i],b[,i])
# Initial values for the parameters
    ct0 = list(epsilon = 0.001, maxit = 50, trace = F)
    alfa1 = coef(glm(zeroy~x.l, family = binomial(link = "logit"),
        na.action = na.omit, control = ct0))
    nb = glm.nb(y~x.l, data=data, link = log,
        na.action = na.omit, control = ct0)
    beta1 = coef(nb)
    k0 = 1/nb$theta
    yu = rep(0,m)
    va = rep(0,m)
    sigu1 = 0.5
    sigv1 = 0.5
    rho = 0
```

```
    glm.logit = wreml.logit(zeroy, x.l, R, alfa1, yu, sigu1)
    alfa = as.vector(glm.logit$alfa)
    sigu = glm.logit$sig1
    u = glm.logit$yu
    p0 = exp(W%*%alfa + R%*%u)/(1+exp(W%*%alfa + R%*%u))
    glm.tnb = wreml.tnb(y, zeroy, x.l, R, beta1, va, sigv1, k0, p0)
    beta = as.vector(glm.tnb$beta)
    sigv = glm.tnb$sig2
    v = glm.tnb$va
    k = glm.tnb$k
    r = t(H) %*% c(u,v)
    coef = c(alfa,beta,r)
# Loop control
    flag = 0; epsilon = 0.001; itmax = 1000
# Begin of it loop
    for (it in 1:itmax)
    {
    IA = getIA(sigu,sigv,rho,m)
# Begin of iter loop
    for (iter in 1:1000)
        {
        theta = as.vector(exp(W %*% alfa + R %*% u))
        p = theta/(1+theta)
        mu = as.vector(exp(X %*% beta + R %*% v))
        f.V = biv.hessen(y,r,H,alfa,u,beta,v,IA,W,X,R,k)
        f = f.V$f
        V = f.V$V
        IV = solve(V)
        coef0 = coef + IV %*% f
        test = try(max(abs((coef-coef0))) < epsilon)
```

```
        if(test) {flag = 1;break}
        coef = coef0
        alfa = coef[1:p1]; beta = coef[(p1+1):(p1+p2)]; r = coef[(p1+p2+1):(p1+p2+2*m)]
        r0 = H %*% r; u = r0[1:m]; v = r0[(m+1):(2*m)]
        }
# End of iter loop
    if(!flag) break; flag = 0
# Get L1, L2, L3
    S = IV[(p1+p2+1):(p1+p2+2*m), (p1+p2+1):(p1+p2+2*m)]
    x = S + r %*% t(r)
    L = L1L2L3(x,m)
    L1 = L$L1; L2 = L$L2; L3 = L$L3
# Get sigu, sigv, rho
    vect = sigu.sigv.rho.v1b(sigu,sigv,rho,L1,L2,L3,m)
    sigu0 = vect[1];sigv0 = vect[2];rho0 = vect[3]
# Get k
    k0 = agetk.ml(y, mu, zeroy, p)
    if(max(abs(c((sigu-sigu0),(sigv-sigv0),(rho-rho0),(k-k0)))) < epsilon)
    {flag = 1; break}
    sigu = sigu0; sigv = sigv0; rho = rho0;k = k0
    }
# End of it loop
    if(!flag) result = list(NULL)
    else
# Get standard errors for the parameters
    {
    se = sqrt(diag(IV))
    se.alfa = se[1:p1]
    se.beta = se[(p1+1):(p1+p2)]
# Get standard error for the k
```

```
    c = max(y)
    t = k/(k+mu)
    f0 = table(y[y>0])
    f = rep(0, c)
    f[as.numeric(names(f0))] = f0
    tot = sum(f0)
    f = tot + f - cumsum(f)
    i = sum(f/(k+1:c-1))
    ii = - sum(f/(k+1:c-1)^2)
    B = log(t)+(1-t)
    B1 = (1-t)^2/k
    ep1 = t^k
    ep2 = (1-ep1)^2
    w33 = (-sum((1-zeroy)*(B1*(1-ep1)+B^2*ep1)/ep2+(1-zeroy)*y*(t/k)^2))-ii
    se.k = sqrt(1/w33)
# Calculate the loglikelihood
    ksi = W%*%alfa + R%*%u; eta = X%*%beta + R%*%v;
    p = exp(ksi)/(1+exp(ksi)); mu = exp(eta)
    loglik = sum(zeroy*log(p) + (1-zeroy)*log(1-p) + (1-zeroy)*(log(gamma(y+k)/gamma(y+1)
    /gamma(k))+k*log(t)+y*log(1-t)-log(1-t^k)))
    result = list(alfa=alfa,beta=beta,sigu=sigu,sigv=sigv,rho=rho,k=k,se.alfa=se.alfa,
    se.beta=se.beta,se.k=se.k,loglik=loglik, u=u,v=v)
    }
# Output
result
}


    # Import the EDCAP data in R
    data = read.table("C://edcap_3112.txt", header=T, as.is=T)
# Zero LOS for inpatients was counted as 1 day
```

```
    for (i in 1:(length(data[,8])))
    {
    if(data[i,8] == 0.5) data[i,8] = 1
    }
# Keep only low risk PSI patients (N=1,877)
    data = subset(data, psihigh == 0)
    attach(data)
    names(data)
    psi.2 = as.numeric(1<psiclass & psiclass<3)
    psi.3 = as.numeric(psiclass>2)
    trt.2 = as.numeric(1<trt & trt<3)
    trt.3 = as.numeric(trt>2)
    edcap = cbind(los, site, trt.2, trt.3, psi.2, psi.3)
    Model = REnbH(edcap)
```

APPENDIX C

ML ESTIMATION SAS CODE IN THE CORRELATED RANDOM EFFECTS POISSON HURDLE MODEL

We will only present the SAS code for the EDCAP study as a simple example. Before running the following SAS code, we need to save the SAS data file, "sas_edcap", in the work library.

Zero LOS for inpatients was counted as 1 day

```
DATA edcap;
SET work.sas_edcap;
IF los = 0.5 THEN los=1;
RUN;
```

Keep only low risk PSI patients (N=1,877)

```
DATA edcapsub;
SET edcap;
one = 1;
y = los;
PSI2 = (psiclass = 2);
PSI3 = (psiclass = 3);
WHERE psihigh eq 0;
```

```
RUN;
DATA sig2;
LENGTH parameter $15.;
INPUT parameter $ estimate;
DATALINES;
sigu 0.5
sigv 0.5
rho 0
;
RUN;
DATA edcap_logit;
SET edcapsub;
IF y>0 THEN y=1;
RUN;
PROC LOGISTIC DATA = edcap_logit ;
MODEL y (event='0') = psi2 psi3 trt2 trt3 /TECH=NEWTON ;
ODS OUTPUT ParameterEstimates=para_logit(RENAME=(VARIABLE=parameter)) ;
RUN;
DATA para_logit(KEEP=parameter estimate);
LENGTH parameter $15.;
SET para_logit;
FORMAT parameter $15.;
IF parameter="Intercept" THEN parameter="cons_inf";
IF parameter="PSI2" THEN parameter="psi2_inf";
IF parameter="PSI3" THEN parameter="psi3_inf";
IF parameter="TRT2" THEN parameter="trt2_inf";
IF parameter="TRT3" THEN parameter="trt3_inf";
RUN;
DATA edcap_p;
SET edcapsub;
```

```
WHERE y^=0;
RUN;
PROC GENMOD DATA = edcap_p;
MODEL y = psi2 psi3 trt2 trt3 /LINK=log DIST=poisson;
ODS OUTPUT ParameterEstimates=para_p;
RUN;
DATA para_p (KEEP=parameter estimate);
LENGTH parameter $15.;
SET para_p;
FORMAT parameter $15.;
IF parameter="Intercept" THEN parameter="cons_p";
IF parameter="PSI2" THEN parameter="psi2_p";
IF parameter="PSI3" THEN parameter="psi3_p";
IF parameter="TRT2" THEN parameter="trt2_p";
IF parameter="TRT3" THEN parameter="trt3_p";
RUN;
DATA para_p;
SET para_p;
IF parameter ^= "Scale";
RUN;
DATA para1;
SET sig2 para_logit para_p;
RUN;
PROC NLMIXED DATA=edcapsub TECH=NEWRAP ABSXTOL=0.001 MAXITER=1000
QTOL=0.001 ABSCONV=0.001 ABSFCONV=0.001 GCONV=1E-7 ABSGCONV=0.001;
BOUNDS sigu > 0 ,sigv > 0, -1 <= rho <= 1;
PARMS/ DATA = para1;
eta1 = psi2_inf*psi2 + psi3_inf*psi3 + trt2_inf*trt2 + trt3_inf*trt3 + cons_inf*one + u;
p0_inflate = exp(eta1) / (1 + exp(eta1));
eta2 = psi2_p*psi2 + psi3_p*psi3 + trt2_p*trt2 + trt3_p*trt3 + cons_p*one + v;
```

```
expeta2 = exp(eta2);
IF y = 0 THEN ll = log(p0_inflate);
ELSE ll = log(1 - p0_inflate) - expeta2 + y*eta2 - lgamma(y + 1) - log(1 - exp(- expeta2));
MODEL y ~ general(ll);
RANDOM u v ~ normal([0,0],[sigu*sigu, sigu*sigu*rho, sigv*sigv]) SUBJECT = site;
PREDICT expeta2 OUT = mu_hurdlep1 (KEEP = stnum pred y trt psiclass RENAME =
(pred = mu_hurdlep));
PREDICT p0_inflate OUT = mu_hurdlep2 (KEEP = stnum pred RENAME = (pred =
p0_hurdlep));
RUN;
```

APPENDIX D

ML ESTIMATION SAS CODE IN THE CORRELATED RANDOM EFFECTS NEGATIVE BINOMIAL HURDLE MODEL

We will only present the SAS code for the EDCAP study as a simple example. Before running the following SAS code, we need to save the SAS data file, "sas_edcap", in the work library.

Zero LOS for inpatients was counted as 1 day

```
DATA edcap;
SET work.sas_edcap;
IF los = 0.5 THEN los=1;
RUN;
```

Keep only low risk PSI patients (N=1,877)

```
DATA edcapsub;
SET edcap;
one = 1;
y = los;
PSI2 = (psiclass = 2);
PSI3 = (psiclass = 3);
WHERE psihigh eq 0;
```

```
RUN;
DATA sig2;
LENGTH parameter $15.;
INPUT parameter $ estimate;
DATALINES;
sigu 0.5
sigv 0.5
rho 0
;
RUN;
DATA edcap_logit;
SET edcapsub;
IF y>0 THEN y=1;
RUN;
PROC LOGISTIC DATA = edcap_logit ;
MODEL y (event='0') = psi2 psi3 trt2 trt3 /TECH=NEWTON ;
ODS OUTPUT ParameterEstimates=para_logit(RENAME=(VARIABLE=parameter)) ;
RUN;
DATA para_logit(KEEP=parameter estimate);
LENGTH parameter $15.;
SET para_logit;
FORMAT parameter $15.;
IF parameter="Intercept" THEN parameter="cons_inf";
IF parameter="PSI2" THEN parameter="psi2_inf";
IF parameter="PSI3" THEN parameter="psi3_inf";
IF parameter="TRT2" THEN parameter="trt2_inf";
IF parameter="TRT3" THEN parameter="trt3_inf";
RUN;
DATA edcap_nb;
SET edcapsub;
```

```
WHERE yˆ=0;
RUN;
PROC GENMOD DATA = edcap_nb;
MODEL y = psi2 psi3 trt2 trt3 /LINK=log DIST=negbin;
ODS OUTPUT ParameterEstimates=para_nb;
RUN;
DATA para_nb (KEEP=parameter estimate);
LENGTH parameter $15.;
SET para_nb;
FORMAT parameter $15.;
IF parameter="Intercept" THEN parameter="cons_nb";
IF parameter="PSI2" THEN parameter="psi2_nb";
IF parameter="PSI3" THEN parameter="psi3_nb";
IF parameter="TRT2" THEN parameter="trt2_nb";
IF parameter="TRT3" THEN parameter="trt3_nb";
IF parameter="Dispersion" THEN parameter="k";
RUN;
DATA para_nb;
SET para_nb;
IF parameter="k" THEN estimate=1/estimate;
RUN;
DATA para2;
SET sig2 para_logit para_nb;
RUN;
PROC NLMIXED DATA=edcapsub TECH=NEWRAP ABSXTOL=0.001 MAXITER=1000
QTOL=0.001 ABSCONV=0.001 ABSFCONV=0.001 GCONV=1E-7 ABSGCONV=0.001;
BOUNDS sigu > 0 ,sigv > 0, -1<=rho<=1; PARMS/ DATA = para2;
eta1 = psi2_inf*psi2 + psi3_inf*psi3 + trt2_inf*trt2 + trt3_inf*trt3 + cons_inf*one + u;
p0_inflate = exp(eta1) / (1 + exp(eta1));
eta2 = psi2_nb*psi2 + psi3_nb*psi3 + trt2_nb*trt2 + trt3_nb*trt3 + cons_nb*one + v;
```

```
expeta2 = exp(eta2);
IF y = 0 THEN ll = log(p0_inflate);
ELSE ll = log(1 - p0_inflate) + lgamma(y + (k)) - lgamma(y + 1) - lgamma(k) - (y +
(k))*log(1 + 1/k*expeta2) + y*log(1/k*expeta2) - log(1 - (1 + 1/k*expeta2)**(-k));
MODEL y ~ general(ll);
RANDOM u v ~ normal([0,0],[sigu*sigu, rho*sigu*sigv, sigv*sigv]) SUBJECT = site;
PREDICT expeta2 OUT = mu_hurdlenb1 (KEEP = stnum pred y trt psiclass RENAME =
(pred = mu_hurdlenb));
PREDICT p0_inflate OUT = mu_hurdlenb2 (KEEP = stnum pred RENAME = (pred =
p0_hurdlenb));
RUN;
```

BIBLIOGRAPHY

[1] M. Aitkin. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55(1):117–128, 1999.

[2] W. Arulampalam and A.L. Booth. Who gets over the training hurdle? A study of the training experiences of young men and women in Britain. *Journal of Population Economics*, 10(2):197–217, 1997.

[3] D. Bohning, E. Dietz, P. Schlattmann, L. Mendonça, and U. Kirchner. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 162(2):195–209, 1999.

[4] N.E. Breslow and D.G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

[5] K.L. Brown, D.A. Ridout, A.P. Goldman, A. Hoskote, and D.J. Penny. Risk factors for long intensive care unit stay after cardiopulmonary bypass in children. *Critical Care Medicine*, 31(1):28, 2003.

[6] R.B. Cunningham and D.B. Lindenmayer. Modeling count data of rare species: some statistical issues. *Ecology*, 86(5):1135–1142, 2005.

[7] M.L. Dalrymple, I.L. Hudson, and R.P.K. Ford. Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. *Computational Statistics and Data Analysis*, 41(3-4):491–504, 2003.

[8] L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York, 2001.

[9] W.H. Fellner. Sparse matrices, and the estimation of variance components by likelihood methods. *Communications in Statistics-Simulation and Computation*, 16(2):439–463, 1987.

[10] M.J. Fine, H.M. Pratt, D.S. Obrosky, J.R. Lave, L.J. McIntosh, D.E. Singer, C.M. Coley, and W.N. Kapoor. Relation between length of hospital stay and costs of care

for patients with community-acquired pneumonia. *The American Journal of Medicine*, 109(5):378–385, 2000.

[11] P.J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(2):149–192, 1984.

[12] W.H. Greene. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working paper no. EC-94-10, Department of Economics, New York University, 1994.

[13] S. Gurmu. Semi-parametric estimation of hurdle regression models with an application to Medicaid utilization. *Journal of Applied Econometrics*, 12(3):225–242, 1997.

[14] S. Gurmu. Generalized hurdle count data regression models. *Economics Letters*, 58(3):263–268, 1998.

[15] S. Gurmu and P.K. Trivedi. Excess zeros in count models for recreational trips. *Journal of Business & Economic Statistics*, 14(4):469–477, 1996.

[16] D.A. Harville. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338, 1977.

[17] D.C. Heilbron. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, 36(5):531–547, 1994.

[18] D.J. Hsu, R.A. Stone, D.S. Obrosky, D.M. Yealy, T.P. Meehan, J.M. Fine, L.G. Graff, and M.J. Fine. Predictors of Timely Antibiotic Administration for Patients Hospitalized With Community-Acquired Pneumonia From the Cluster-Randomized EDCAP Trial. *The American Journal of the Medical Sciences*, 339(4):307, 2010.

[19] P.M. Kuhnert, T.G. Martin, K. Mengersen, and H.P. Possingham. Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion. *Environmetrics*, 16(7):717–748, 2005.

[20] D. Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

[21] A.H. Lee, K. Wang, KK Yau, G.J. McLachlan, and S.K. Ng. Maternity length of stay modelling by Gamma mixture regression with random effects. *Biometrical Journal*, 49(5):750–764, 2007.

[22] A.H. Lee, K. Wang, K.K.W. Yau, and P.J. Somerford. Truncated negative binomial mixed regression modelling of ischaemic stroke hospitalizations. *Statistics in Medicine*, 22(7):1129–1139, 2003.

[23] A.H. Lee, L. Xiang, and W.K. Fung. Sensitivity of score tests for zero-inflation in count data. *Statistics in Medicine*, 23(17):2757–2769, 2004.

[24] C.E. McCulloch, S.R. Searle, and J.M. Neuhaus. *Generalized, Linear, and Mixed Models*. John Wiley & Sons, Inc., New York, 2001.

[25] C.A. McGilchrist. Estimation in generalized mixed models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):61–69, 1994.

[26] C.A. McGilchrist and K.K.W. Yau. The derivation of BLUP, ML, REML estimation methods for generalised linear mixed models. *Communications in Statistics-Theory and Methods*, 24(12):2963–2980, 1995.

[27] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, Inc., New York, 2001.

[28] Y. Min and A. Agresti. Modeling nonnegative data with clumping at zero: A survey. *Journal of the Iranian Statistical Society*, 1(1-2):7–33, 2002.

[29] Y. Min and A. Agresti. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5(1):1–19, 2005.

[30] J. Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3):341–365, 1986.

[31] M.S. Niederman, J.S. McCombs, A.N. Unger, A. Kumar, and R. Popovian. The cost of treating community-acquired pneumonia. *Clinical Therapeutics*, 20(4):820–837, 1998.

[32] M.J. Page, L.S. Poritz, S.J. Kunselman, and W.A. Koltun. Factors affecting surgical risk in elderly patients with inflammatory bowel disease. *Journal of Gastrointestinal Surgery*, 6(4):606–613, 2002.

[33] I. Pardoe and C.A. Durham. Model Choice Applied to Consumer Preferences. In *Proceedings of the 2003 Joint Statistical Meetings*. Citeseer, 2003.

[34] HD Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545, 1971.

[35] W. Pohlmeier and V. Ulrich. An econometric model of the two-part decisionmaking process in the demand for health care. *Journal of Human Resources*, 30(2):339–361, 1995.

[36] H.E. Quintero, A. Abebe, and D.A. Davis. Zero-Inflated Discrete Statistical Models for Fecundity Data Analysis in Channel Catfish, Ictalurus punctatus. *Journal of the World Aquaculture Society*, 38(2):175–187, 2007.

[37] S.W. Raudenbush, M.L. Yang, and M. Yosef. Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, pages 141–157, 2000.

[38] B. Renaud, A. Santin, E. Coma, N. Camus, D. Van Pelt, J. Hayon, M. Gurgui, E. Roupie, J. Hervé, M.J. Fine, et al. Association between timing of intensive care unit admission and outcomes for emergency department patients with community-acquired pneumonia. *Critical Care Medicine*, 37(11):2867, 2009.

[39] M. Ridout, C.G.B. Demetrio, and J. Hinde. Models for count data with many zeros. In *Proceedings of the XIXth International Biometric Conference*, volume 19, pages 179–192, 1998.

[40] C.E. Rose, S.W. Martin, K.A. Wannemuehler, and B.D. Plikaytis. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics*, 16(4):463–481, 2006.

[41] P. Schlattmann, E. Dietz, and D. Boehning. Covariate adjusted mixture models and disease mapping with the program DismapWin. *Statistics in Medicine*, 15(7-9):919–929, 1996.

[42] J.X. Song. Zero-inflated Poisson regression to analyze lengths of hospital stays adjusting for intra-center correlation. *Communications in Statistics-Simulation and Computation*, 34(1):235–241, 2005.

[43] R. Thompson. Maximum likelihood estimation of variance components. *Statistics*, 11(4):545–561, 1980.

[44] K. Wang, K.K.W. Yau, and A.H. Lee. A hierarchical Poisson mixture regression model to analyse maternity length of hospital stay. *Statistics in Medicine*, 21(23):3639–3654, 2002.

[45] K. Wang, K.K.W. Yau, A.H. Lee, and G.J. McLachlan. Two-component Poisson mixture regression modelling of count data with bivariate random effects. *Mathematical and Computer Modelling*, 46(11-12):1468–1476, 2007.

[46] P. Wang, M.L. Puterman, I. Cockburn, and N. Le. Mixed Poisson regression models with covariate dependent rates. *Biometrics*, 52(2):381–400, 1996.

[47] A.H. Welsh, R.B. Cunningham, C.F. Donnelly, and D.B. Lindenmayer. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, 88(1-3):297–308, 1996.

[48] T. Xie and M. Aickin. A truncated Poisson regression model with applications to occurrence of adenomatous polyps. *Statistics in Medicine*, 16(16):1845–1857, 1997.

[49] K.K.W. Yau, A.H. Lee, and A.S.K. Ng. Finite mixture regression model with random effects: application to neonatal hospital length of stay. *Computational Statistics & Data Analysis*, 41(3-4):359–366, 2003.

[50] D.M. Yealy, T.E. Auble, R.A. Stone, J.R. Lave, T.P. Meehan, L.G. Graff, J.M. Fine, D.S. Obrosky, and S.M. Edick. The emergency department community-acquired pneumonia trial:: Methodology of a quality improvement intervention. *Annals of Emergency Medicine*, 43(6):770–782, 2004.

[51] D.M. Yealy, T.E. Auble, R.A. Stone, J.R. Lave, T.P. Meehan, L.G. Graff, J.M. Fine, D.S. Obrosky, M.K. Mor, J. Whittle, et al. Effect of increasing the intensity of implementing pneumonia guidelines: a randomized, controlled trial. *Annals of Internal Medicine*, 143(12):881, 2005.